Machine Learning for Genomic Medicine

Giorgio Valentini
Dipartimento di Informatica
Università degli Studi di Milano

Computer Science Department

UNIVERSITÀ DEGLI STUDI DI MILANO

Anacleto Lab

Computational Biology and Bioinformatics

Genomic Medicine

“Personalised Medicine refers to a medical model using characterisation of individuals’ phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention.”

Horizon 2020 Advisory Group for Social Challenge “Health, demographic Change and Wellbeing”

Genomic, Precision, Personalized Medicine

- Precision diagnostics: patients stratification on the basis of their biomolecular profiles

- Precision therapeutics: therapies targeted to the biomolecular profiles of patients

- Omics biotechnologies generate heterogeneous big data to profile patients

- Editing technologies able to modify the genome

Goals of Genomic Medicine (GM):

1) Determine how variations in the DNA of individuals can affect the risk of different diseases

2) Find causal explanations so
that targeted therapies can be designed.

Genomic medicine challenges

✓ No well-targeted therapies available for most pathologies

✓ Most of clinically validated targeted therapies are not actually curative:

- only a subset of patients respond to therapies

- only limited sets of bio markers are available

✓ Most monotherapies not able to deal with the multi-pathway involved in most diseases

↓

✓ Need for innovative omics technologies to measure hidden “cell variables”

✓ Need for innovative Artificial Intelligence methods to analyze the data and make inferences

✓ Need for multi-disciplinary teams (Medicine, Biotechnology, Artificial Intelligence, Bioinformatics)

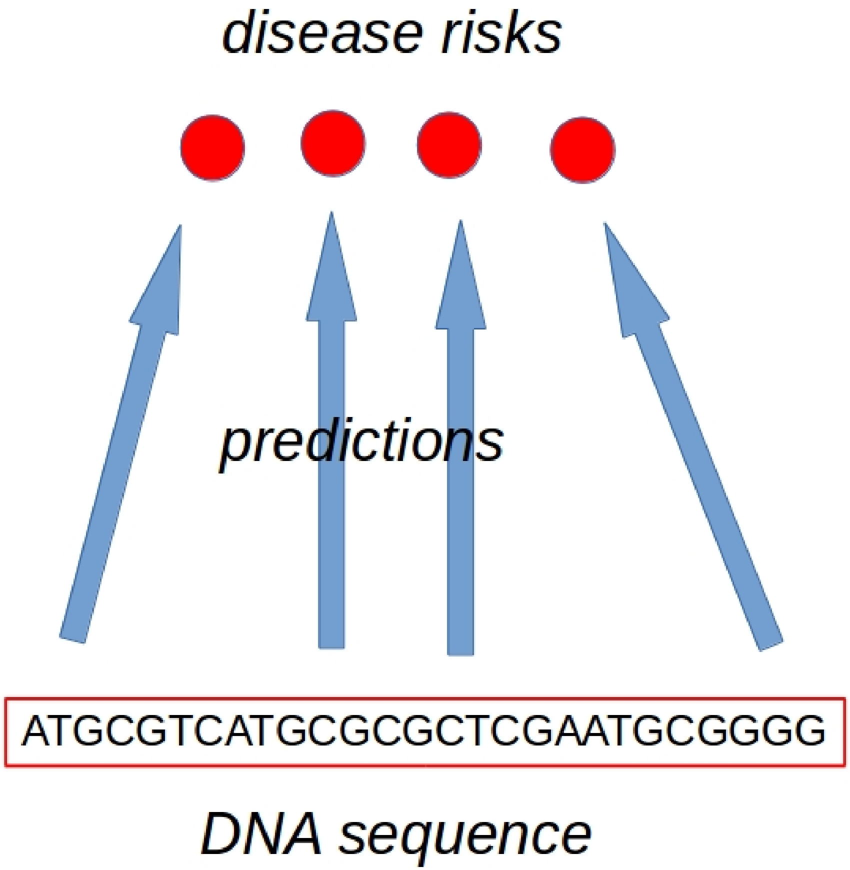
Why we need Machine Learning for GenomicMedicine ?

- The scale and complexity of genomic data dwarfs the small number of measurements that are traditionally used in laboratory tests (Rubin, *Nature*, 2015)

- ML models the relationship between DNA and the quantities of key molecules in the cell (cell variables,) may be associated with disease risks (Leung et al., *Proc. of IEEE* 2016).

- The effects of genetic variation and potential therapies can be explored quickly, cheaply, and more accurately than can be achieved using laboratory experiments and model organisms.

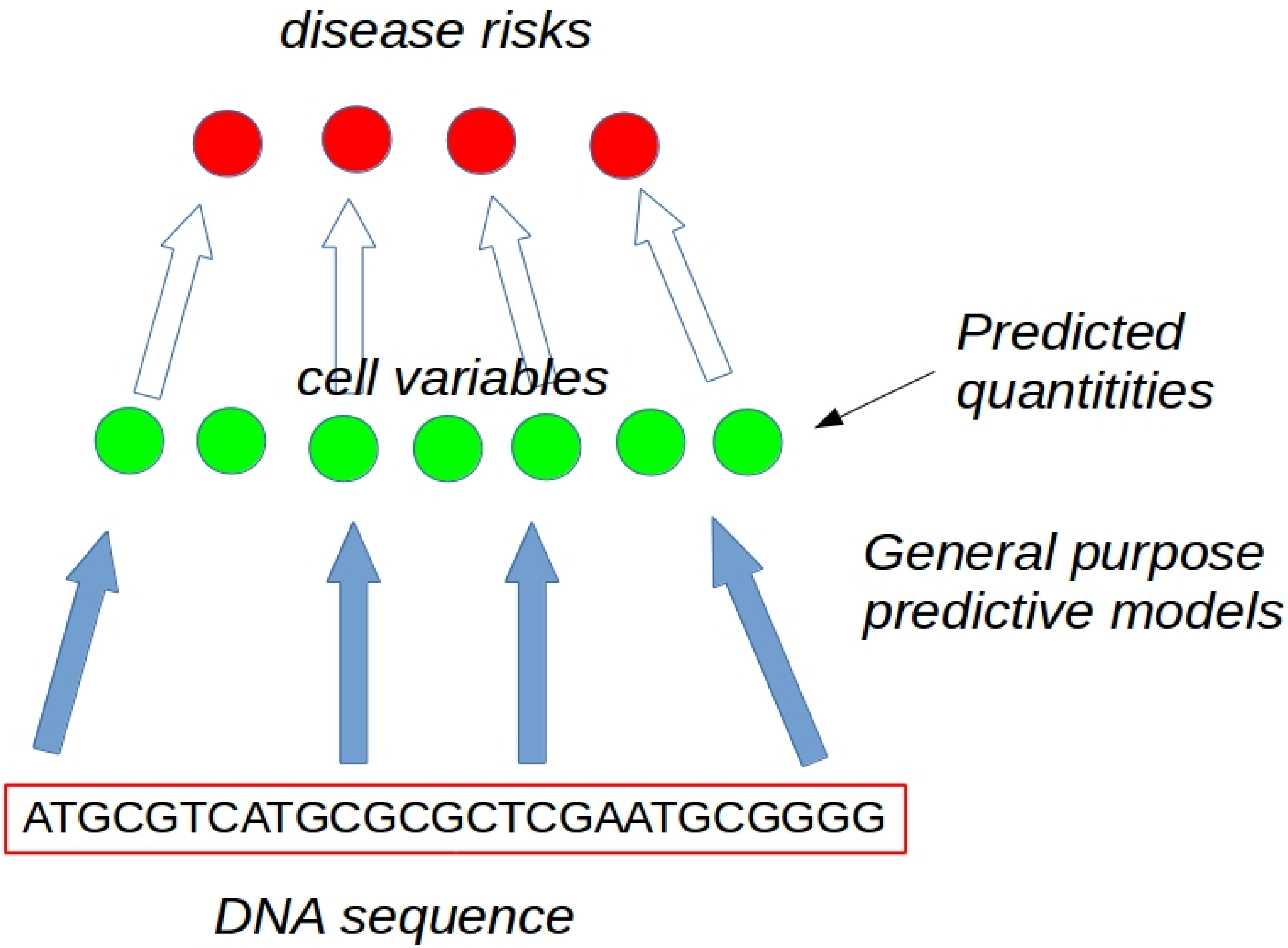
Phenotype from genotype prediction as a ML supervised problem.



Direct inference is very hard

need for hidden variables:
underlying biophysical chemical pathways, interactions, intermediate regulatory machinery

Predicting cell variables (molecular phenotypes) is simpler



Why using cell variables?

1) more directly related to genotypes

2) high throughput technologies generate data profiling cell variables

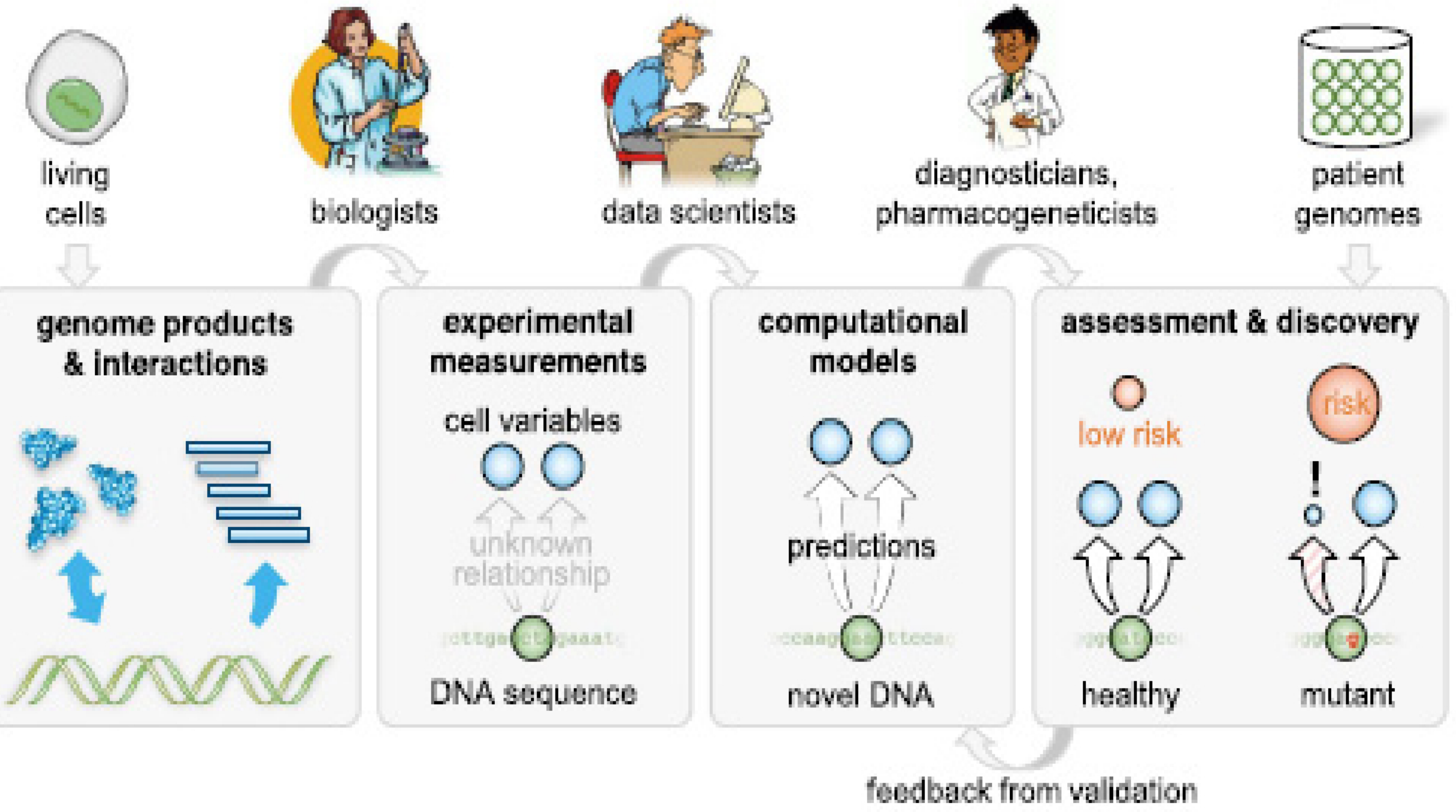
3) cell variables help to discover targets for therapies

Assays to measure cell variables

- DNA microarray
- Universal protein binding microarrays (PBMs)
- ChIP-chip
- High-throughput sequencing technologies:
  — identifying protein binding sites
  — sequencing the genomes of different organisms in evolutionary studies,
  — profiling the genomes of individuals in medical studies for the purpose of discovering variations
  — analysis of transcripts
- DNA methylation
- Assays for chromatin structure,
- Assays for RNA or protein folding
- ...

<u>Wealth of data: must be processed with computational methods</u>

CELL BIOLOGY, MACHINE LEARNING, AND GENOMIC MEDICINE

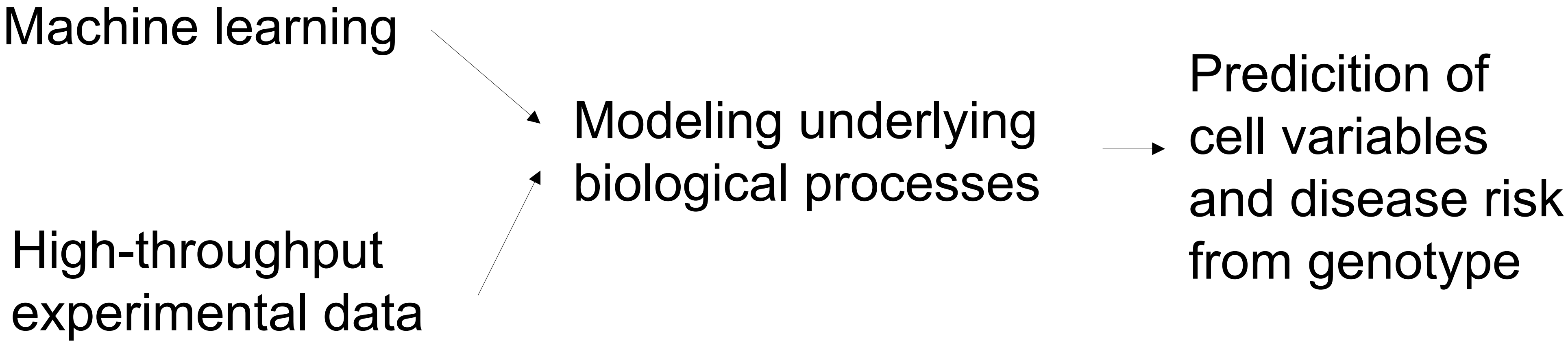


Leung et al, Machine Learning in Genomic Medicine:
A Review of Computational Problems and Data Sets *Proc of IEEE*, 2016

Why ML is necessary for Genomic Medicine?

The details of many interactions, quantities, and processes in the cell are ‘‘hidden’’ from us because we do not have the technology to systematically measure them .

In other words, the few cell variables that we can observe are the outcome of many layers of interacting cell variables that we cannot observe.



Machine learning

High-throughput experimental data

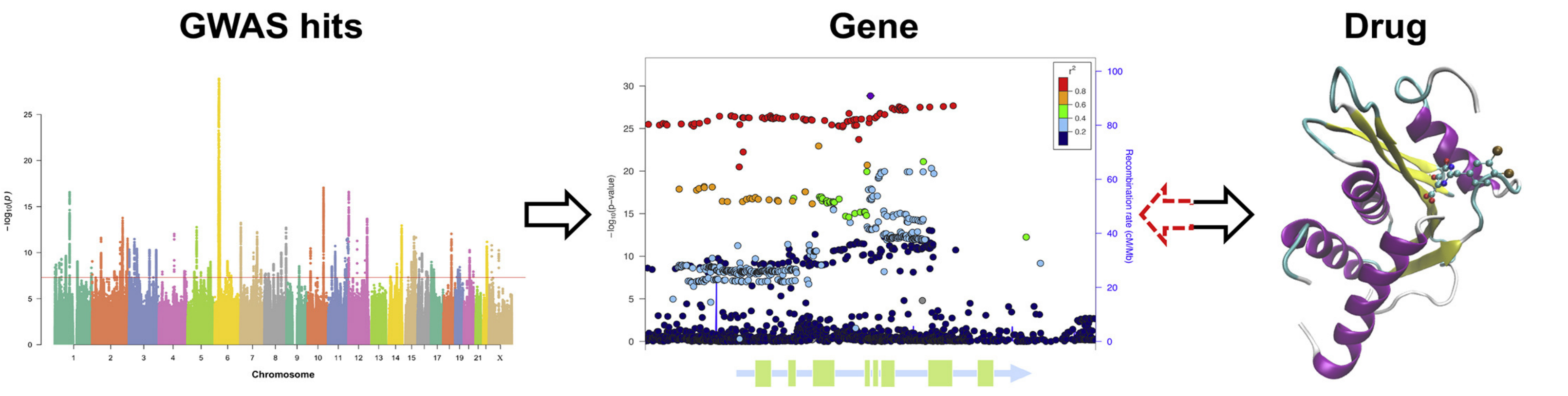
Modeling underlying biological processes

Predicition of cell variables and disease risk from genotype

A paradigmatic example: approaches for mapping genetic variants with disease risks

1) through association (GWAS)
2) through the use of comparative genomics.
3) <u>Through advanced ML methods trained on well-designed experimental data</u>

Genome-Wide Association Studies

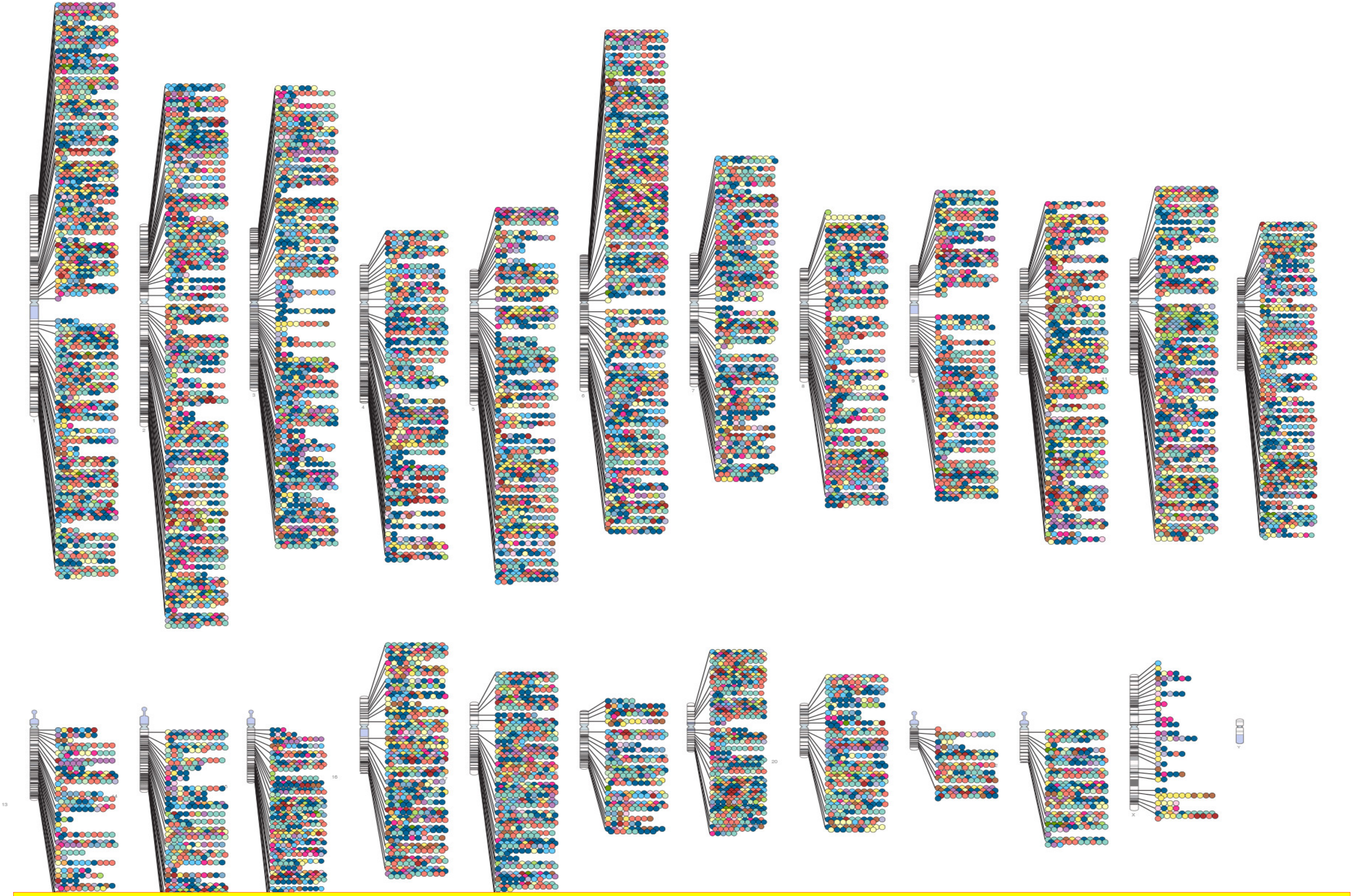


Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

GWAS detect how traits within a population can be related to variants in particular genomic locations using microarray and sequencing techniques.

P. M. Visscheret al. 10 Years of GWAS Discovery: Biology, Function, and Translation, *Amer. J. Human Genetics*, Vol. 101, Issue 1, 2017

Genome-Wide Association Studies



SNP-trait associations with p-value<5.0 X 10^{-8} in the GWAS Catalog (NHGRI-EBI)

Drawbacks of GWAS

- Difficult to establish a statistical significance between a potentially causal variant with a change in risk for particular disease
- Indicates correlation, not causation.
- GWAS provides a huge number of putative causal mutations → researchers biased toward candidates that have greater “narrative potential”
- Assessing the statistical significance of an immense number of SNPs is challenging and requires careful multiple-hypothesis correction.

Methods based on Evolutionary Conservation:

Mostly rely on sequence conservation.

Rationale behind sequence conservation:

A. Evolution driven by two forces:

- the slow accumulation of random mutations
- selective pressures against mutations that damage fitness within a population.

B. Genomes compared across species: sequence conservation is the effect of selective pressure (if time enough is passed)

Conservation scores are available for multiple organisms:

a) *phastCons* (Siepel et al 2005), *GERP* (Cooper et al. 2005) *phyloP* (Pollard et al. 2010). Conservation scores for each position in the human genome can be viewed online.

Methods based on Evolutionary Conservation:

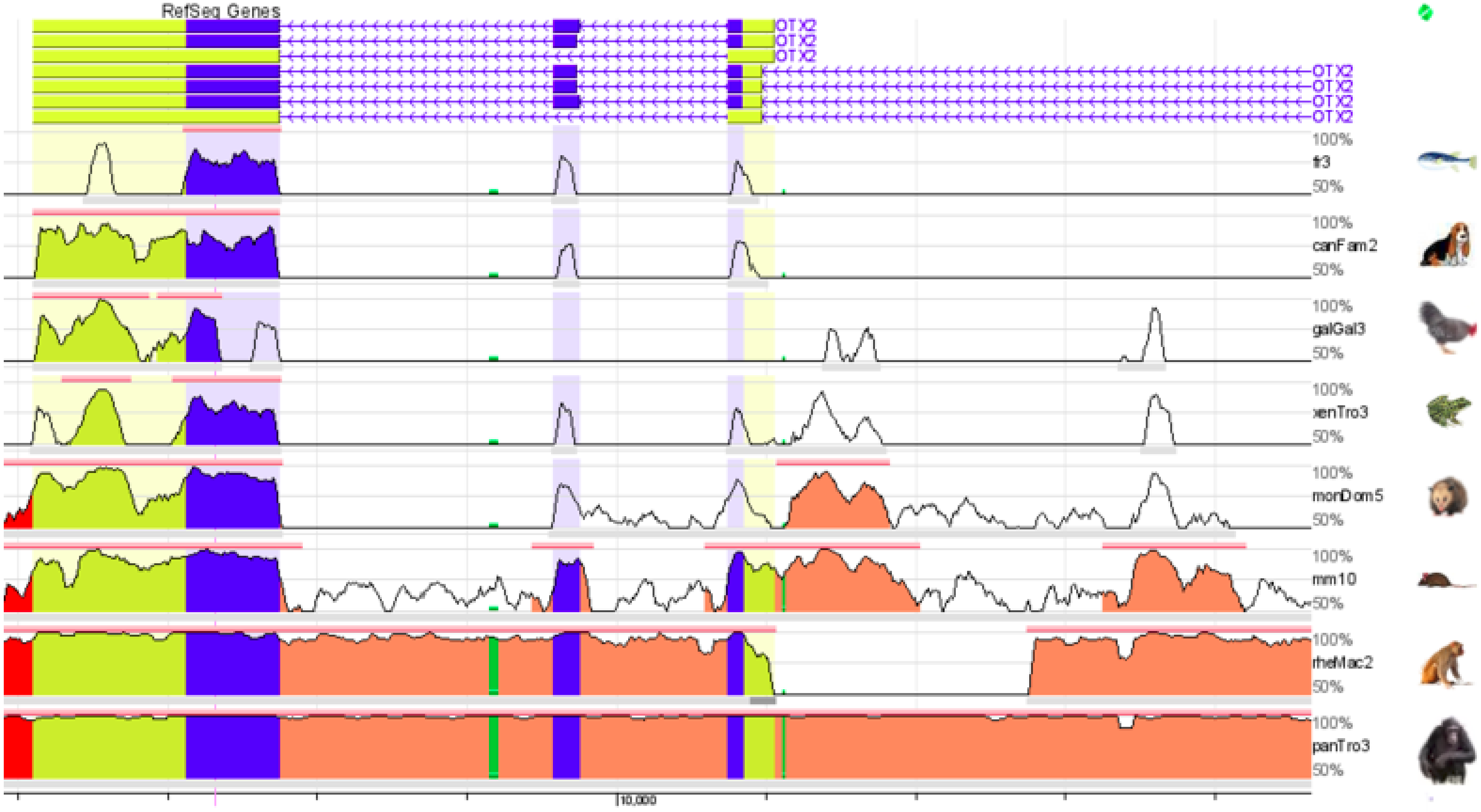


image from the ECR browser

OTX2: the encoded protein acts as a transcription factor and plays a role in brain and sensory organ development

Deleterious and pathogenic mutations

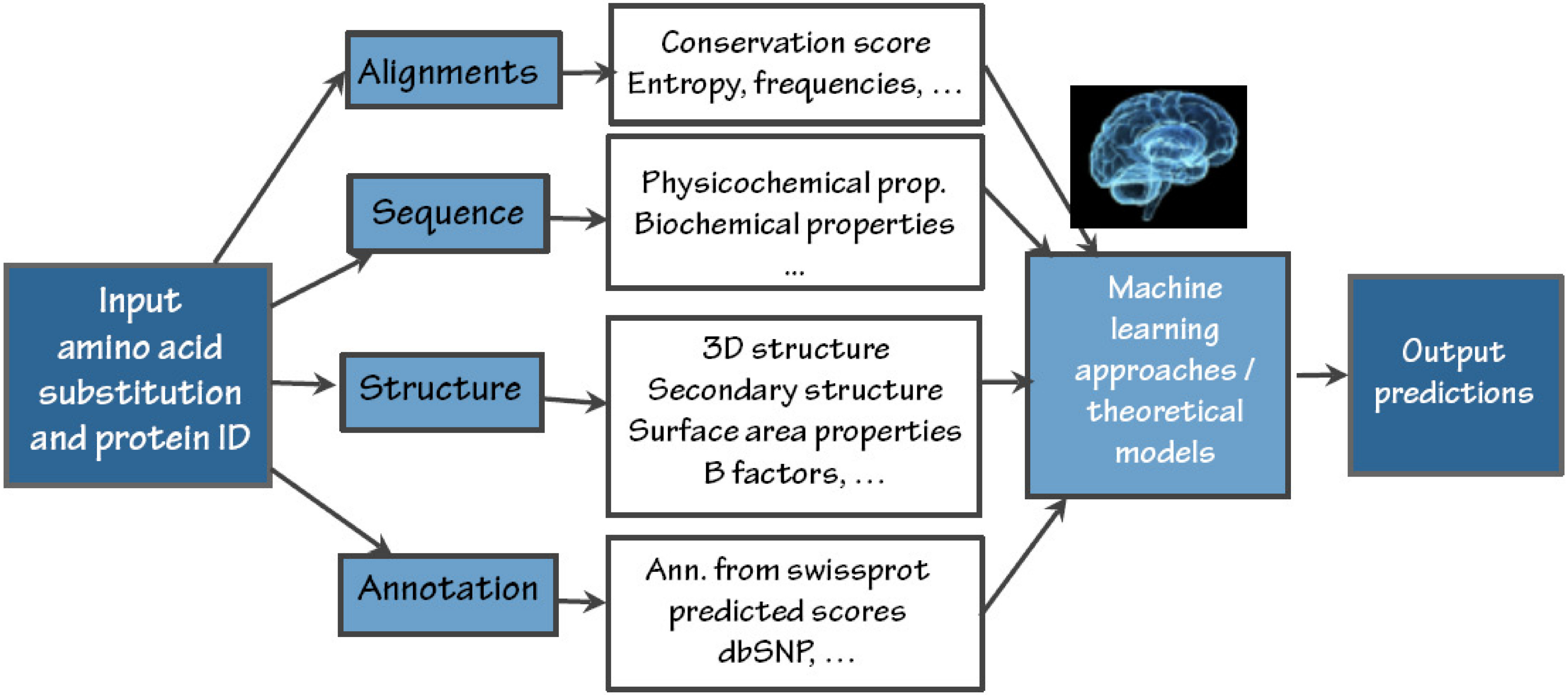
- Mutation that lowers reproductive fitness is called *deleterious*

- Mutation that causes a disease is called *pathogenic* (MacArthur et al. *Nature*, 2014).

- Conservation only provides information about deleteriousness, but deleteriousness is related to pathogenicity

Identification of deleterious variants: first proposed methods relied on coding sequences

Typical pipeline for identification of deleterious variants found in coding sequence (WES, panels, …)



Input amino acid substitution and protein ID

Alignments

Sequence

Structure

Annotation

Conservation score
Entropy, frequencies, …

Physicochemical prop.
Biochemical properties
…

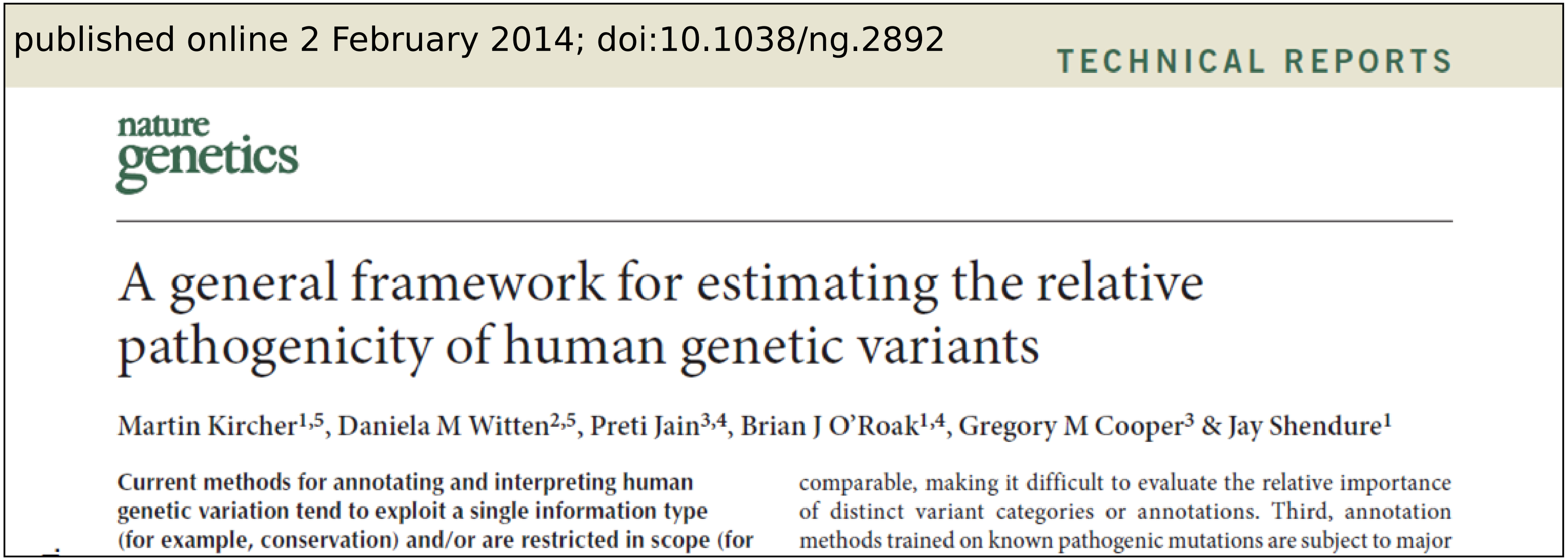
3D structure
Secondary structure
Surface area properties
B factors, …

Ann. from swissprot
predicted scores
dbSNP, …

Machine learning approaches / theoretical models

Output predictions

Combined Annotation Dependent Depletion (CADD) scores



published online 2 February 2014; doi:10.1038/ng.2892

TECHNICAL REPORTS

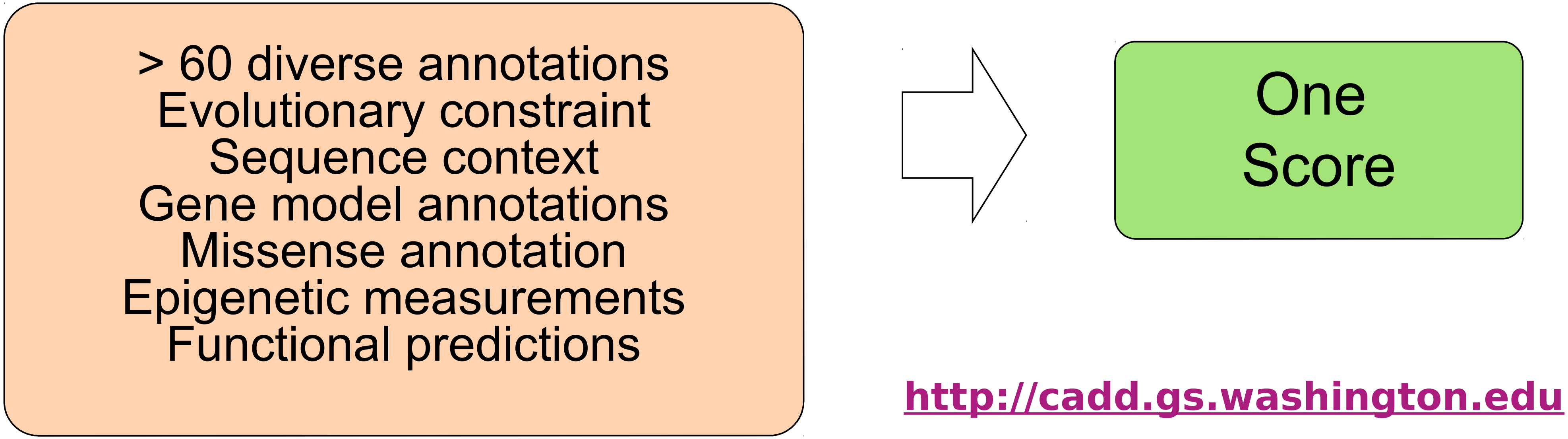
nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for

comparable, making it difficult to evaluate the relative importance of distinct variant categories or annotations. Third, annotation methods trained on known pathogenic mutations are subject to major

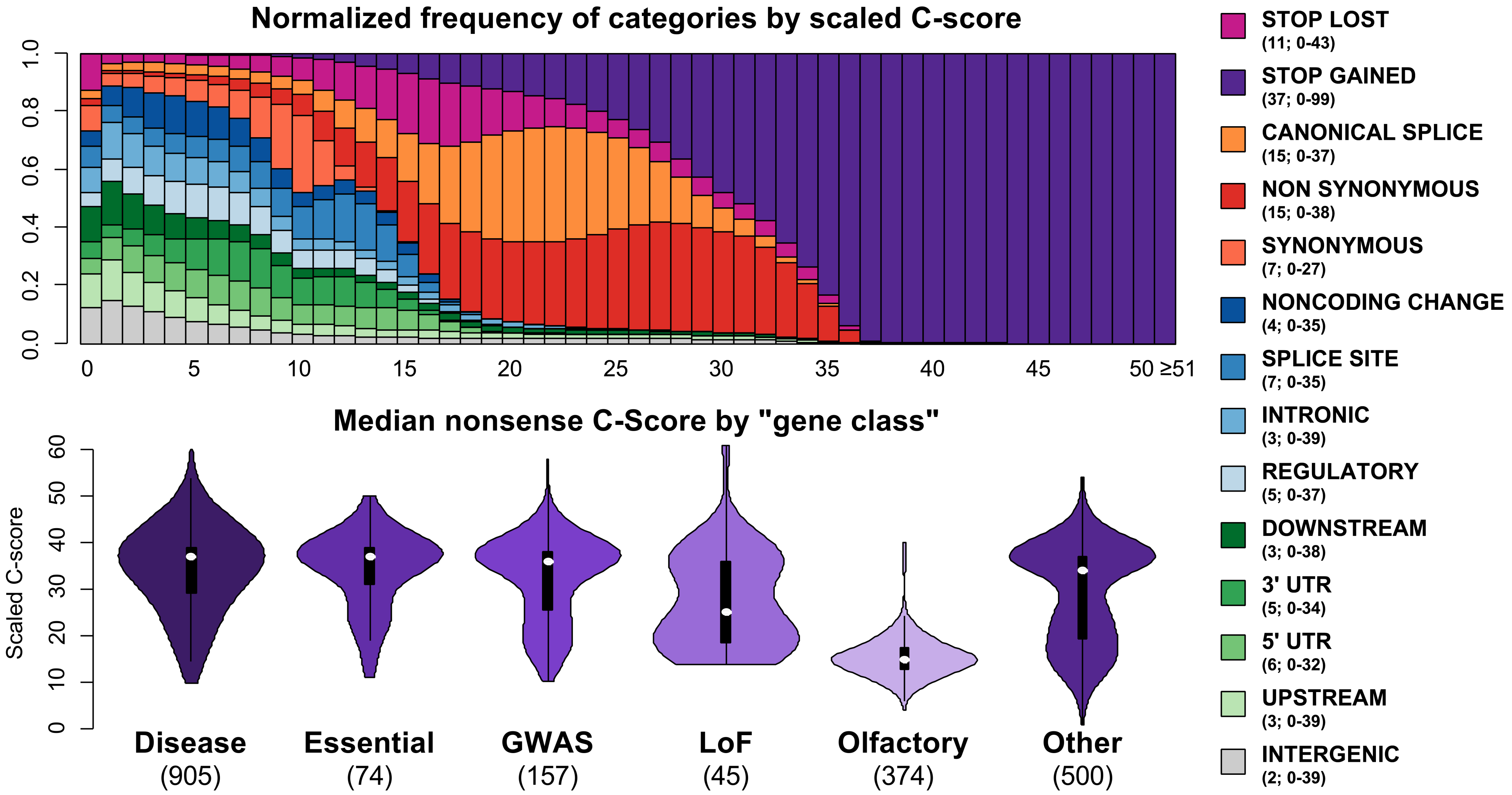


> 60 diverse annotations
Evolutionary constraint
Sequence context
Gene model annotations
Missense annotation
Epigenetic measurements
Functional predictions

One Score

http://cadd.gs.washington.edu

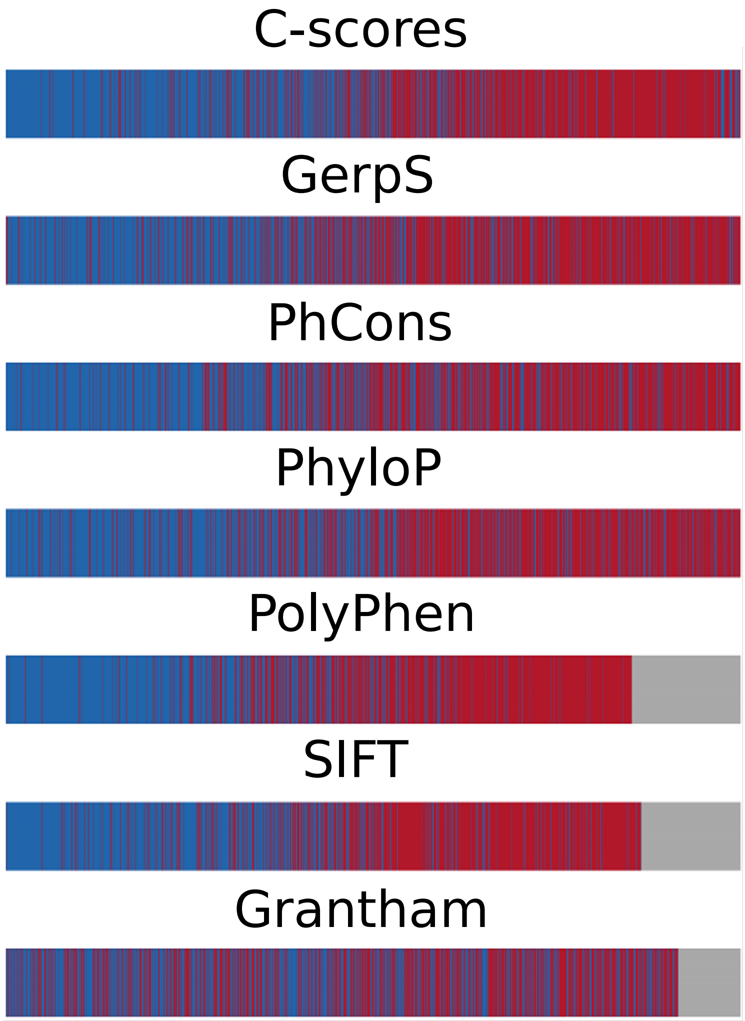
Scoring all 8.6 x 10^9 possible SNVs

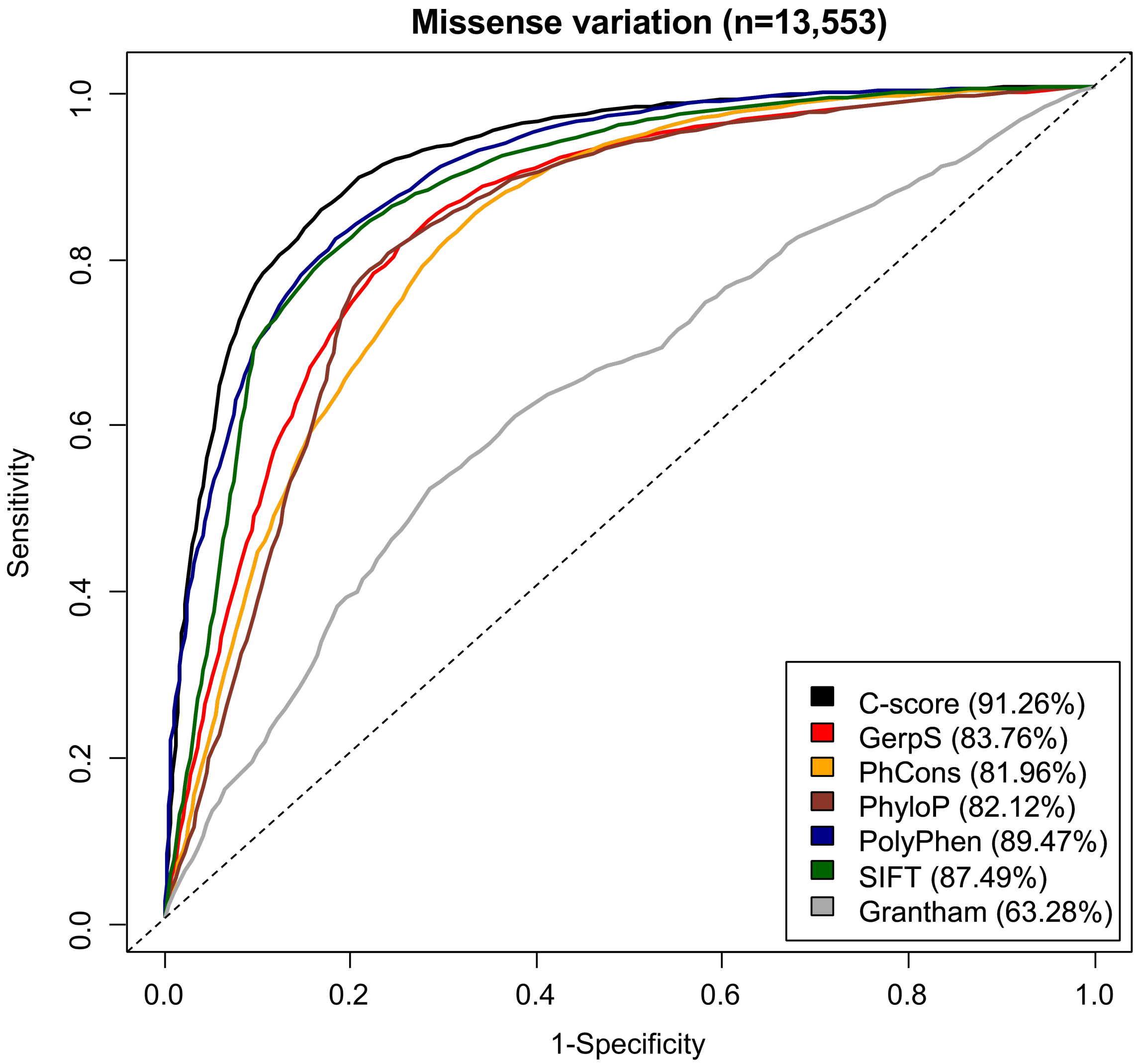

Normalized frequency of categories by scaled C-score

Median nonsense C-Score by "gene class"

"PHRED"-like scaling: $-10 \cdot \log_{10}(\text{rank}/(8.6 \cdot 10^9))$

Separating ClinVar pathogenic from ESP benign sites (AF > 5%)





Measuring "deleteriousness" as proxy for pathogenicity

Proxy benign

~15 million fixed or nearly fixed **human-derived alleles** (*i.e.* 95-100% derived allele frequency)

vs.

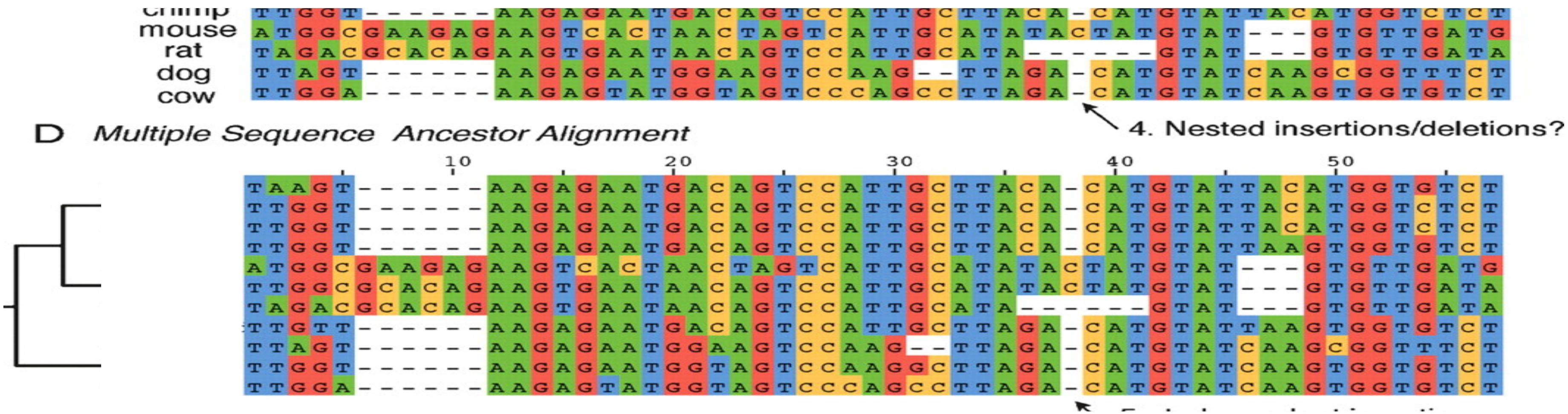
Proxy deleterious

~15 million **simulated mutations** (empirical model of primate sequence evolution)

Fixed/nearly fixed human-derived alleles

- Ensembl Enredo-Pecan-Ortheus (EPO) six primate alignments to obtain the ancestral sequence A
- Include human reference genome sites that:
  - differ from A
  - with AF < 5% (1000G project)
  - Low frequency derived variants (DAF <95%) excluded

Nearly fixed human derived alleles (likely to be benign)


chimp
mouse
rat
dog
cow
4. Nested insertions/deletions?
D Multiple Sequence Ancestor Alignment
10
20
30
40
50

Modified from Paten B et al. Genome Res. 2008;18:1829-1843

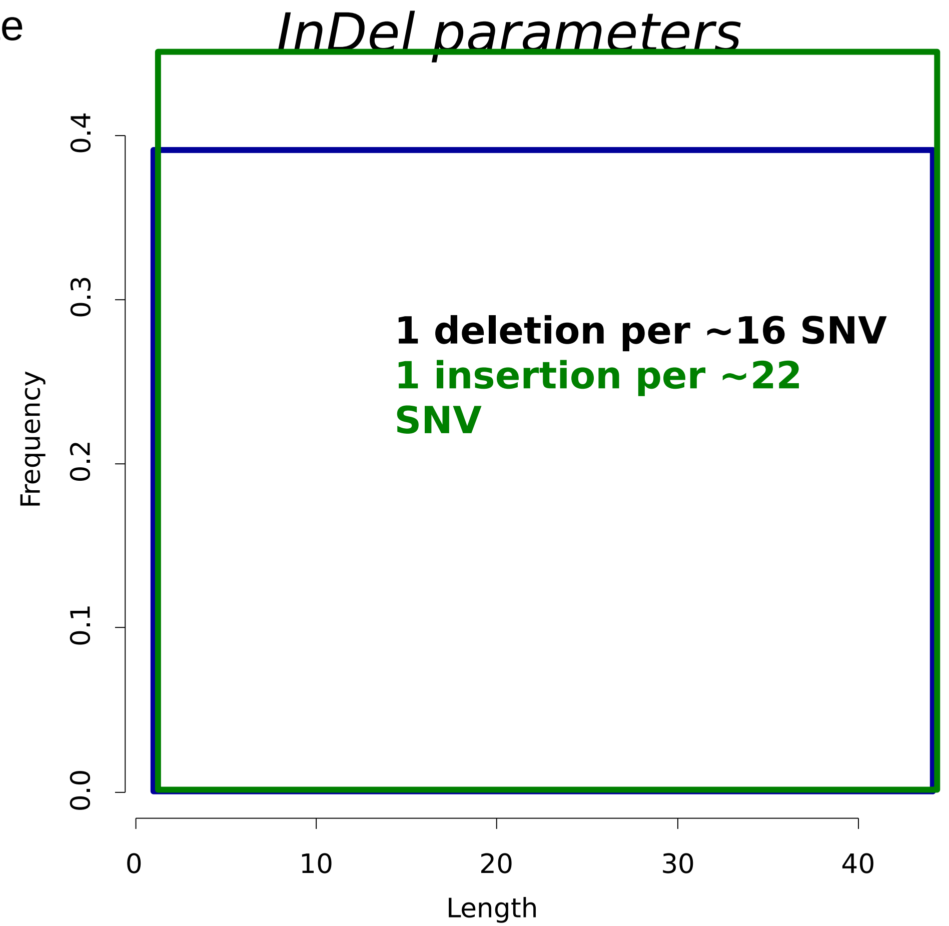
Simulation of variants

Substitution parameters

Local mutation rate → µ

CpG island mutation rate → µCpG

µ	0.522%	µCpG	4.979%
		%CpG	1.9%
A>C	4.574%		
A>G	17.507%		
A>T	4.042%		
C>A	4.610%	C>A	3.00%
C>G	4.350%	C>G	2.48%
C>T	14.896%	C>T	44.60%
G>A	14.978%	G>A	44.38%
G>C	4.358%	G>C	2.51%
G>T	4.620%	G>T	3.03%
T>A	4.001%		
T>C	17.469%		
T>G	4.595%		


InDel parameters

1 deletion per ~16 SNV

1 insertion per ~22 SNV

- Rates obtained by comparison between the human reference genome and and the inferred ancestral human-chimpanzee sequence
- Local mutation rate (µ) as determined from 1.1Mb windows across the genome (±5 x 100 Kb blocks) neighborhood

Simulation of variants

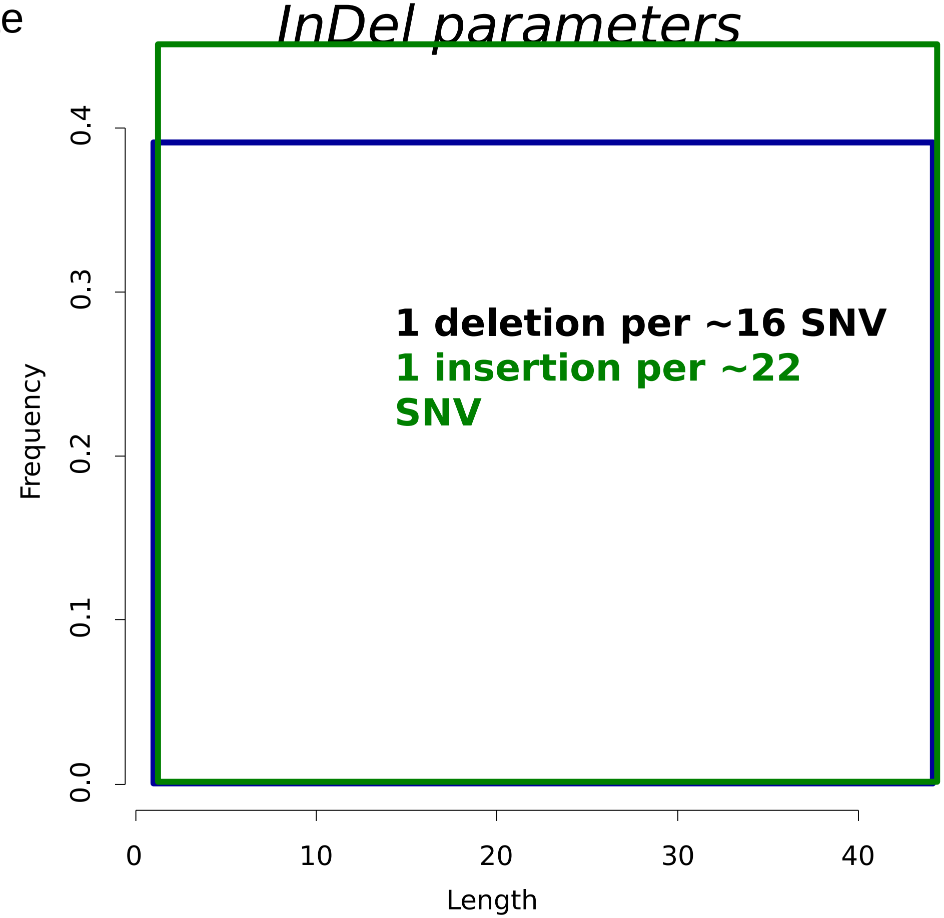
Substitution parameters

Local mutation rate

CpG island mutation rate

μ	**0.522%**	**μCpG**	**4.979%**
		%CpG	**1.9%**
A>C	**4.574%**		
A>G	**17.507%**		
A>T	**4.042%**		
C>A	**4.610%**	**C>A**	**3.00%**
C>G	**4.350%**	**C>G**	**2.48%**
C>T	**14.896%**	**C>T**	**44.60%**
G>A	**14.978%**	**G>A**	**44.38%**
G>C	**4.358%**	**G>C**	**2.51%**
G>T	**4.620%**	**G>T**	**3.03%**
T>A	**4.001%**		
T>C	**17.469%**		
T>G	**4.595%**		

InDel parameters



1 deletion per ~16 SNV
1 insertion per ~22 SNV

- Rates obtained by comparison between the human reference genome and and the inferred ancestral human-chimpanzee sequence
- Local mutation rate (μ) as determined from 1.1Mb windows across the genome (±5 x 100 Kb blocks) neighborhood

CADD annotations in more detail

- **Genome-wide measurements that correlate with function/biological constraint :**
  - Accessibility of chromatin (DNase, FAIRE-seq, ...)
  - Activity of region (polyA-transcript expression, histone marks)
  - Predicted overlapping transcription factor motifs
  - Segway genomic segmentation type inferred from ENCODE data
  - Conservation scores: GERP + human-free Phast and PhyloP

+ Gene model based information, e.g. Ensembl VEP:

- Type of change, amino acids, position in transcript, ...
- Distance to transcription start/end and splice sites
- Grantham, PolyPhen, SIFT scores

Variant Effect Predictor:
https://www.ensembl.org/info/docs/tools/vep

CADD 1.0 uses a linear SVM as classifier

Rows = variants (~30M)

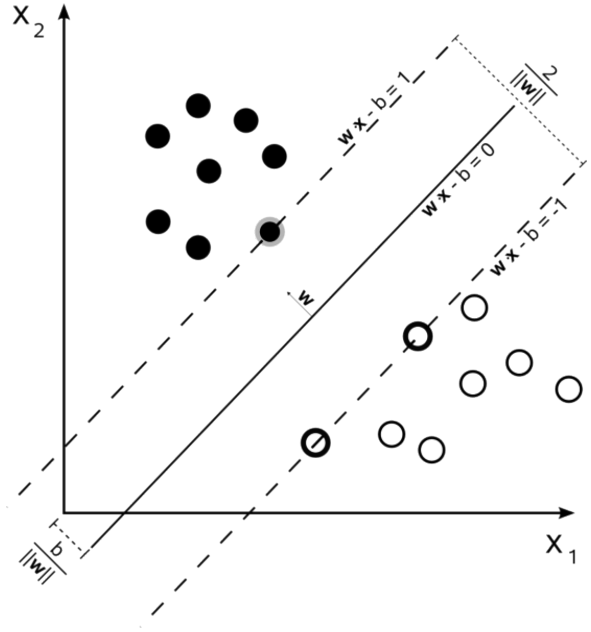
y=0 for proxy benign

y=1 for proxy deleterious

Columns = annotations

$X_1,\ldots,X_n$ 63 annotations, indicator variables, and subset of interactions

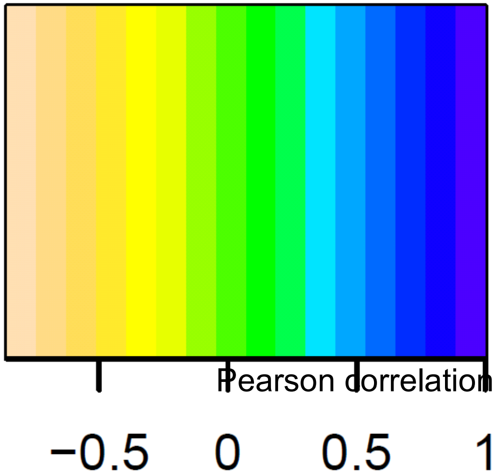
Linear SVM

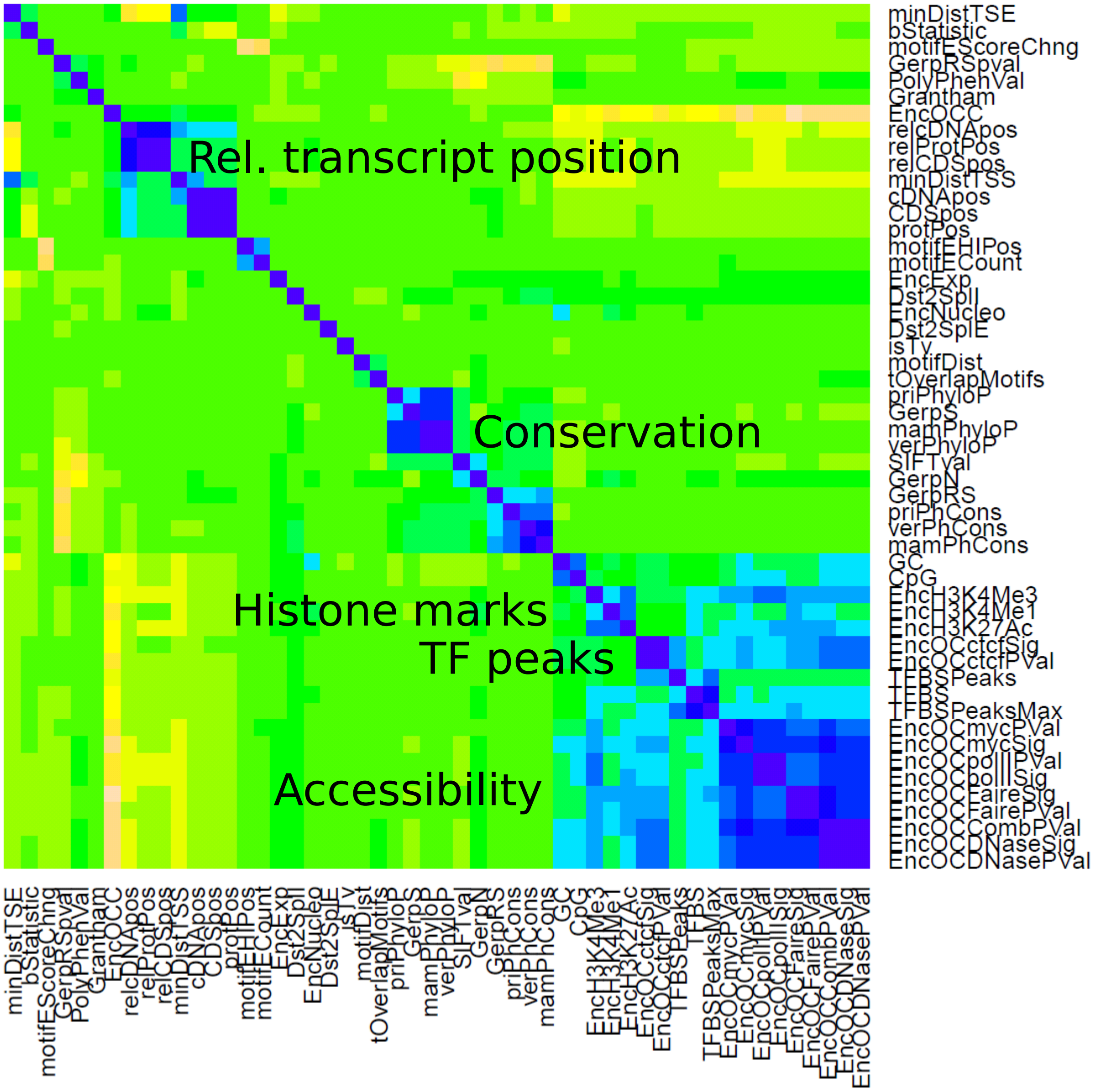


Challenge 1: Dimensionality & correlations

High dimensionality (>900 features, >400 due to amino acid replacements) and correlated features

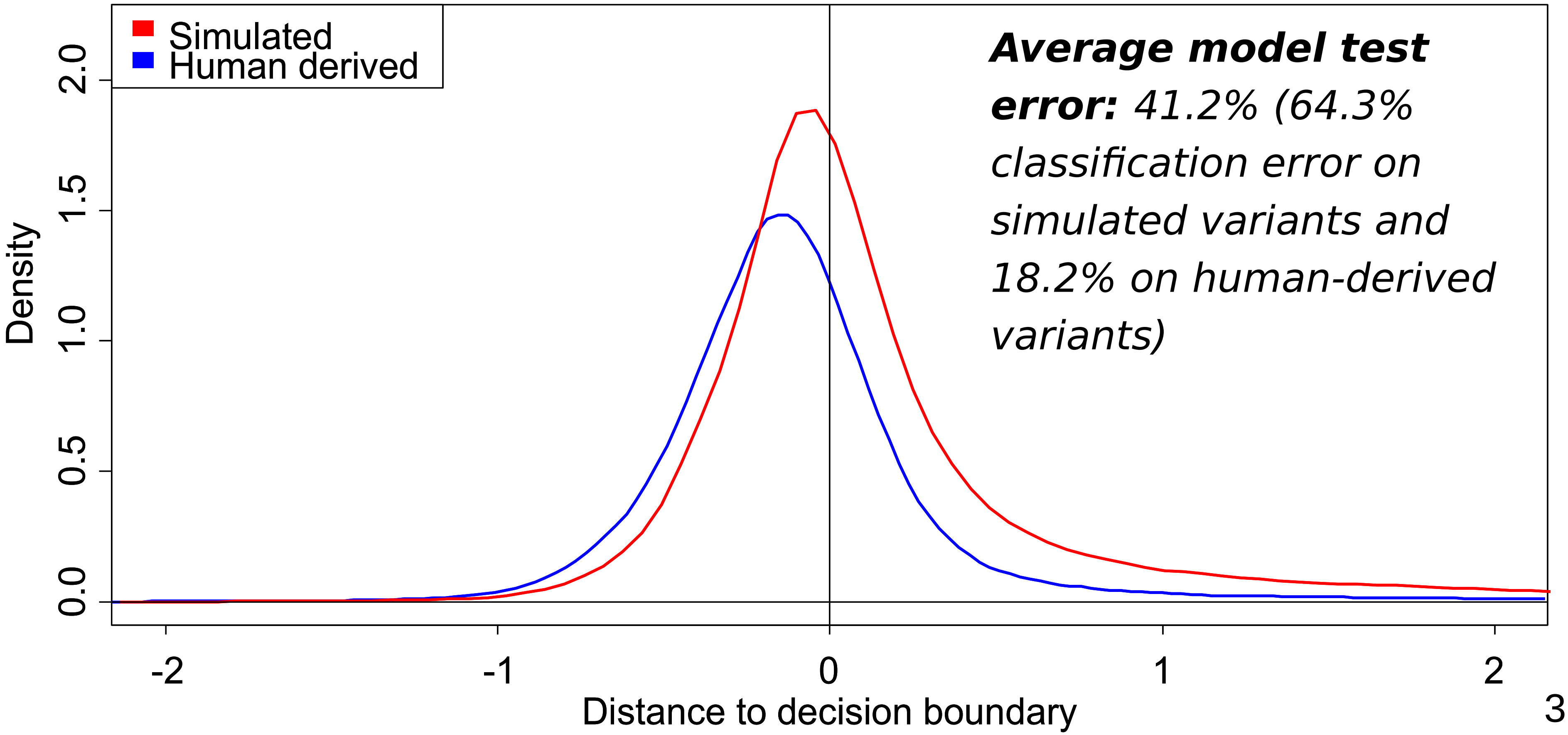
Sparse and structured





Challenge 2: Large amount of mislabeling

By definition, large proportion (< ~80%) of simulated and small proportion (< ~5%) of human-derived variants expected to be incorrectly labeled



***Average model test error:** 41.2% (64.3% classification error on simulated variants and 18.2% on human-derived variants)*

Challenge 3: Model choice

Choosing model and training parameters

Selection of interactions terms /

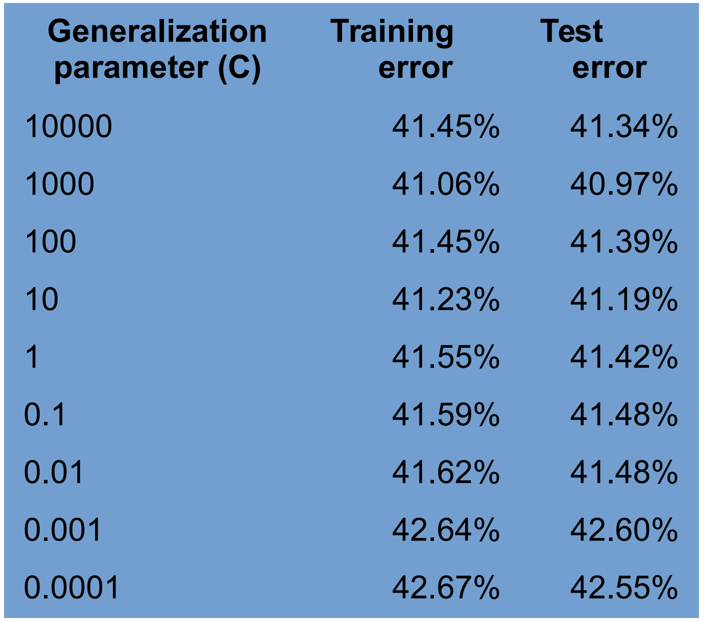
- non-linear models
- Additional model parameters, e.g.
- class weights, regularization constant
- C for SVM, L1/L2 penalty for a logistic
- regression
- Training termination criteria

Training computationally expensive

E.g. >200G for training matrix in R

Training and evaluation objectives different

CADD v1.0 SVM training



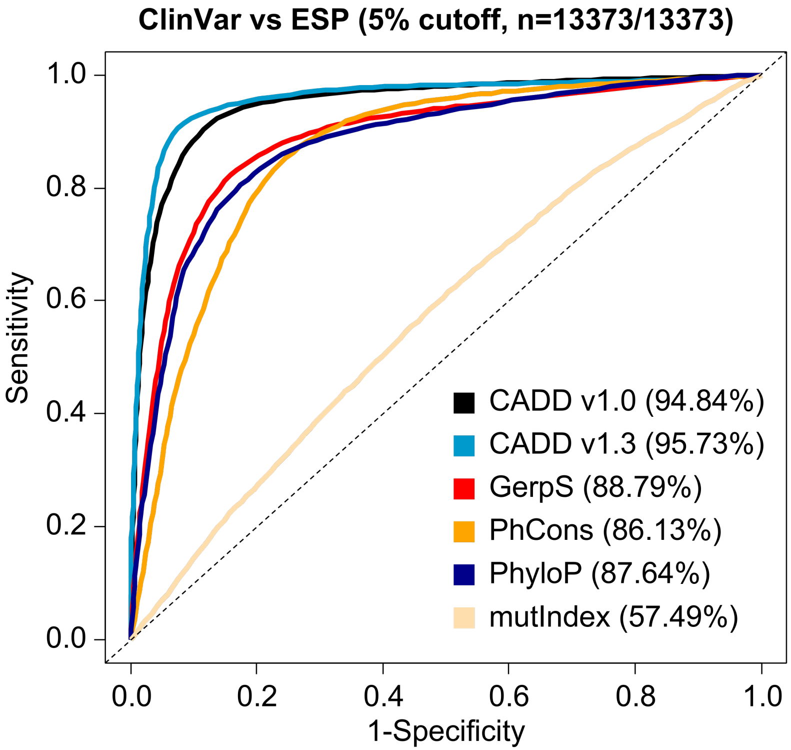
Generalization parameter (C)	Training error	Test error
10000	41.45%	41.34%
1000	41.06%	40.97%
100	41.45%	41.39%
10	41.23%	41.19%
1	41.55%	41.42%
0.1	41.59%	41.48%
0.01	41.62%	41.48%
0.001	42.64%	42.60%
0.0001	42.67%	42.55%

LIBOCAS <20G of memory, no convergence within a week of computation

CADD v1.0: 10 runs sampling matching number of simulated variants. Averaging model coefficients after 24h of training

Latest release: CADD v1.3 (July 2015)

- **Improved training data**:
- Updated Ensembl EPO whole
  - genome alignments
  - Increased number of
  - **training variants (+5%)**

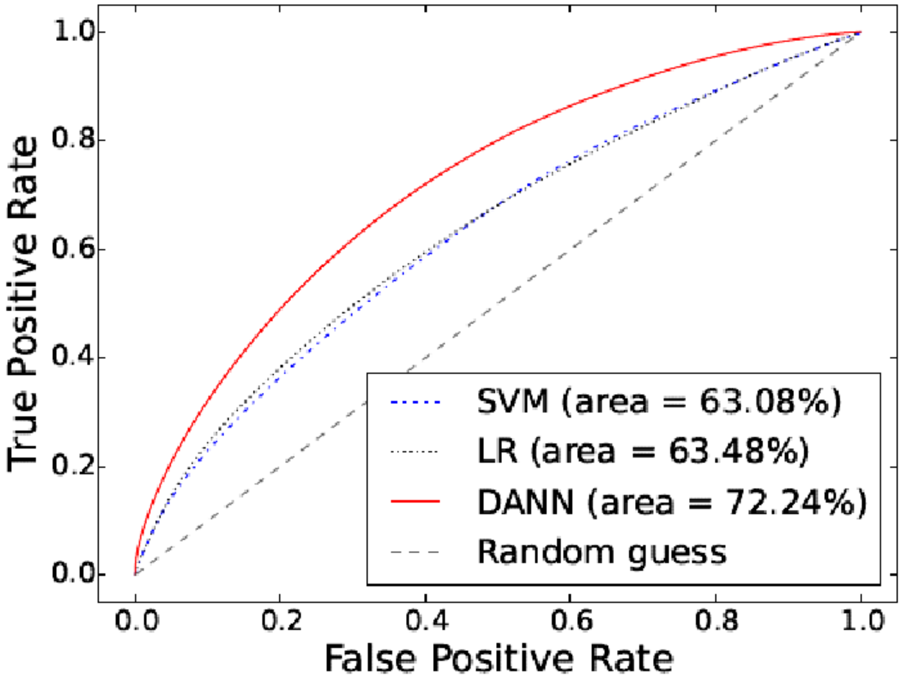


- GraphLab 1.4 (Guestrin, 2016) logistic regression model trained in **11.1 min**

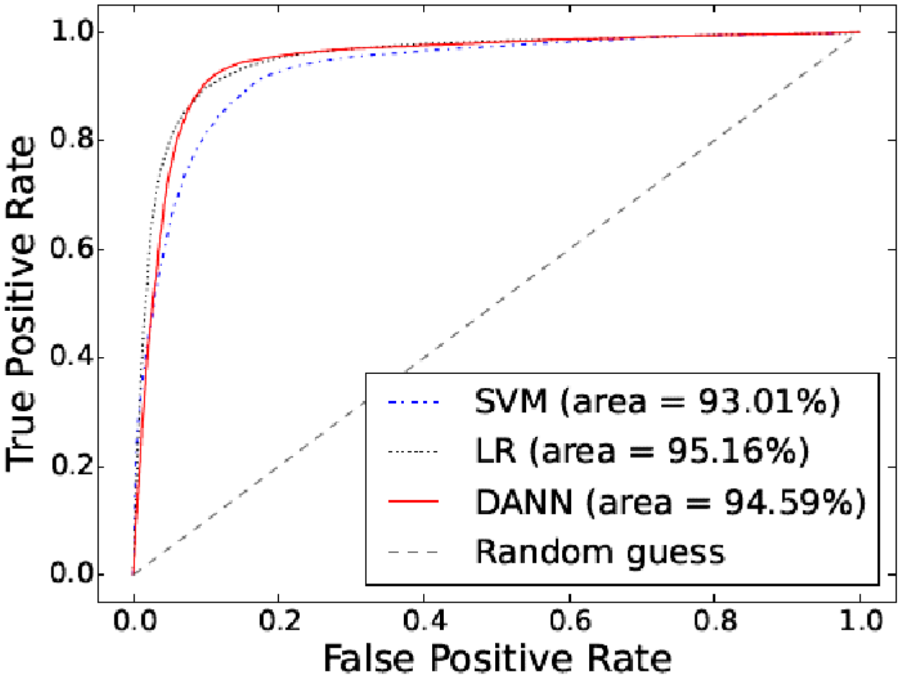
Can deep learning improve results?

Quang et al. used CADD data set to train a deep neural network to improve performances (Quang et al. *Bioinformatics*, 2014)

Training objective (test data)



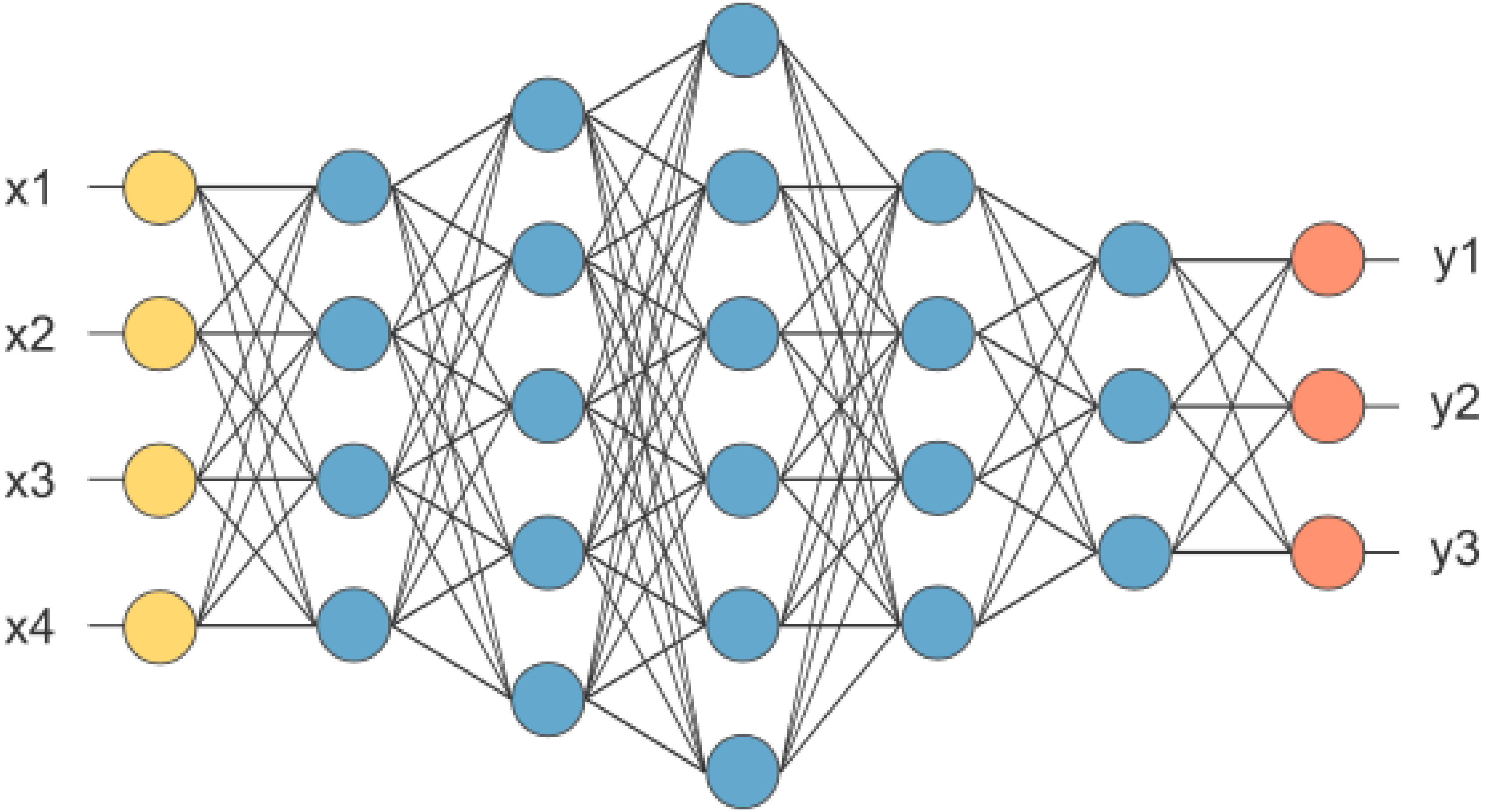
ClinVar vs ESP



Deep learning showed promising results in several contexts of Genomic Medicine:

- Feedforward neural networks for alternative splicing patterns (Leung et al. *Bioinformatics* 2014)
- Convolutional neural networks for binding specificity by Alipanahi et al. *Nat. Biotechnol.* 2015)
- Convolutional neural networks for chromatin effects prediction (Zhou and Troyanskaya, *Nat. Methods*, 2015)
- Deep autoencoder to predict survival in Liver Cancer (Chaudury et al. *Clinical Cancer Research* 2018*)*

Deep learning



INPUT

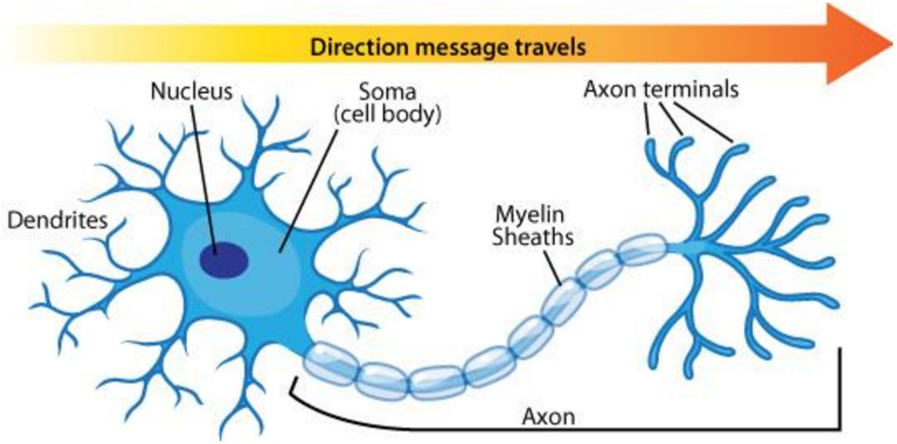
OUTPUT

HIDDEN NEURONS

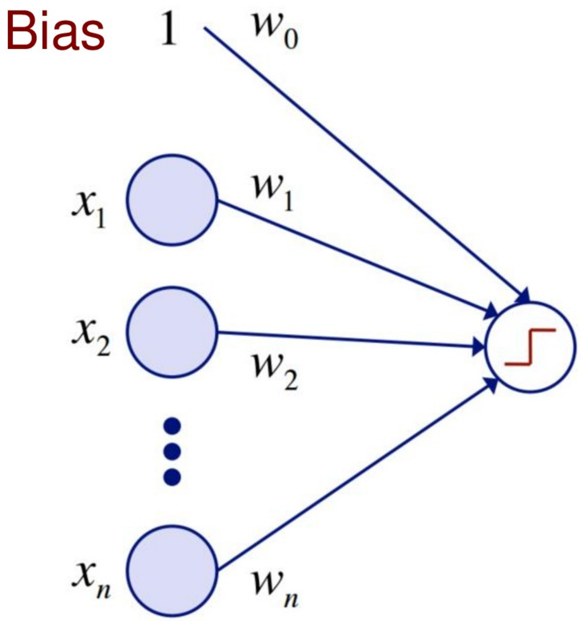
Image taken from www.opennn.net/

Perceptron

- Inspired by neuron



- Simple binary classifier
  - Linear decision boundary



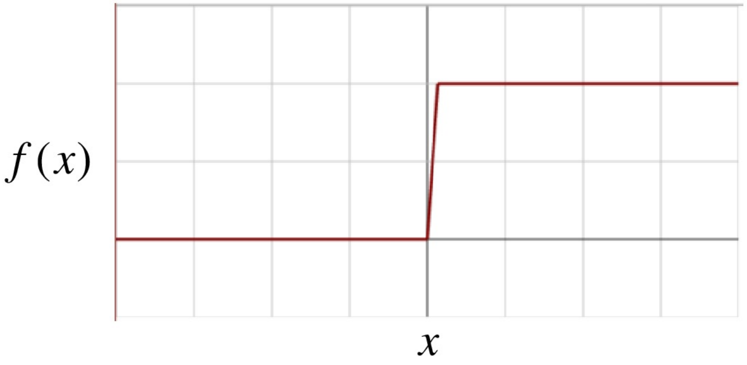
$$o = \begin{cases} 1 \ \text{if} \ w_0 + \sum_{i=1}^{n} w_i x_i > 0 \\ 0 \ \text{otherwise} \end{cases}$$

$$x_{1,A} \ldots x_{1000,T}$$

Activation function

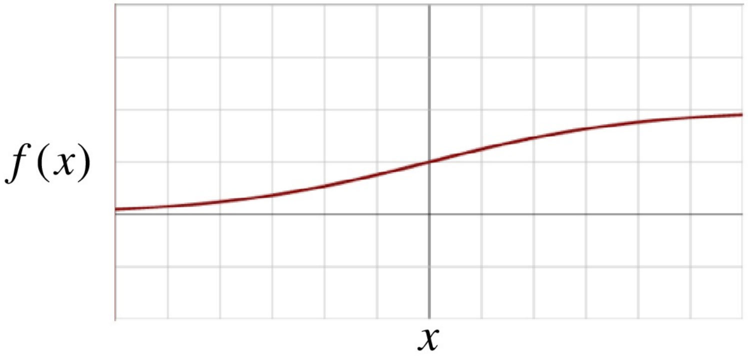
- What makes the neuron "fire"?
  - Step function

$$f(x)=\begin{cases}0 & \text{if } x<0\\ 1 & \text{if } x\geq 0\end{cases}$$



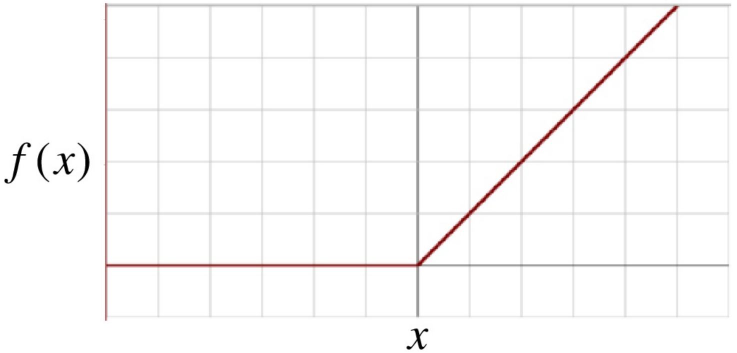
  - Sigmoid function

$$f(x)=\frac{1}{1+e^{-x}}$$



  - Rectified linear unit (ReLU)

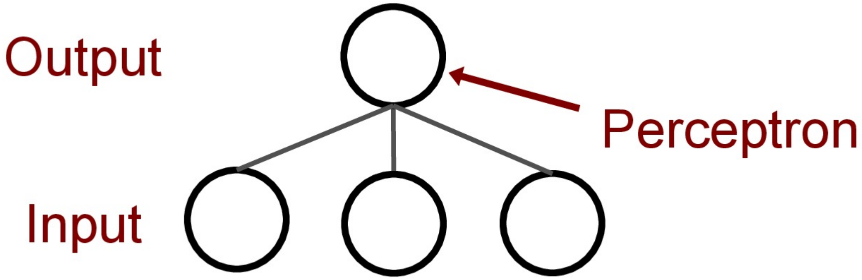
$$f(x)=\max(0,x)$$



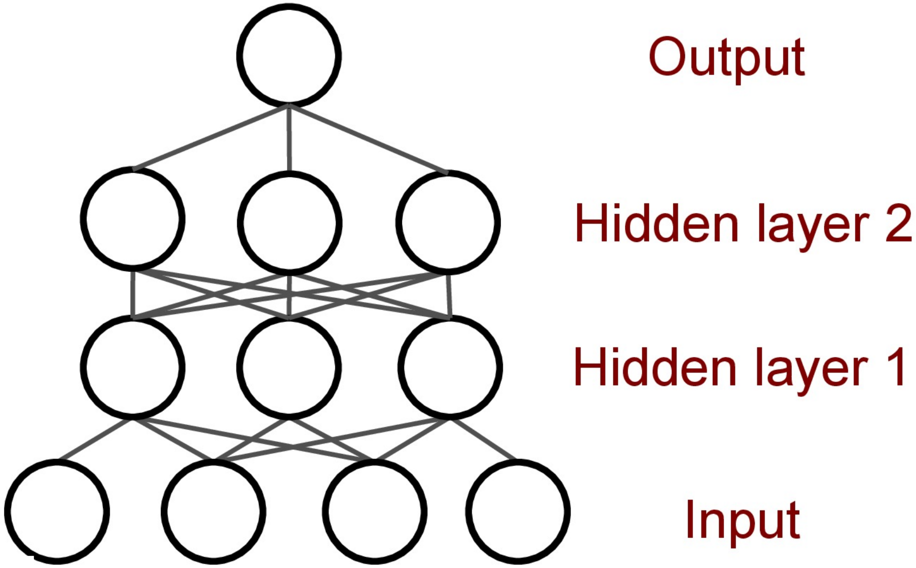
Images from Wikipedia: Activation function

Neural networks

- Single perceptron not useful in practice



- Neural network combines layers of perceptrons
- Learn "hidden" features
- Complex decision boundary
- Train with backpropagation

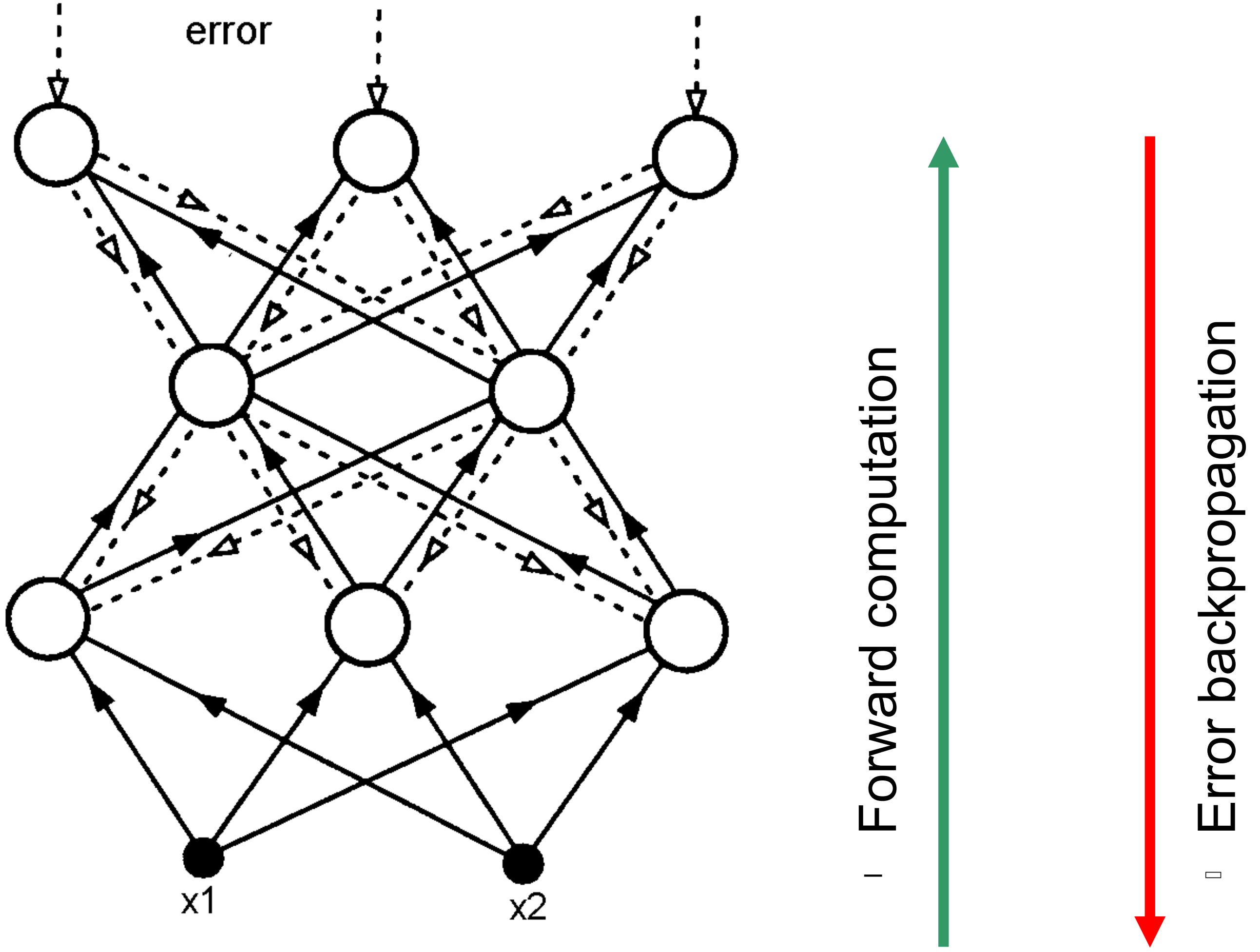


A schematic view of the error backprogation algorithm for a 3-layers neural network

Output

Hidden
layers

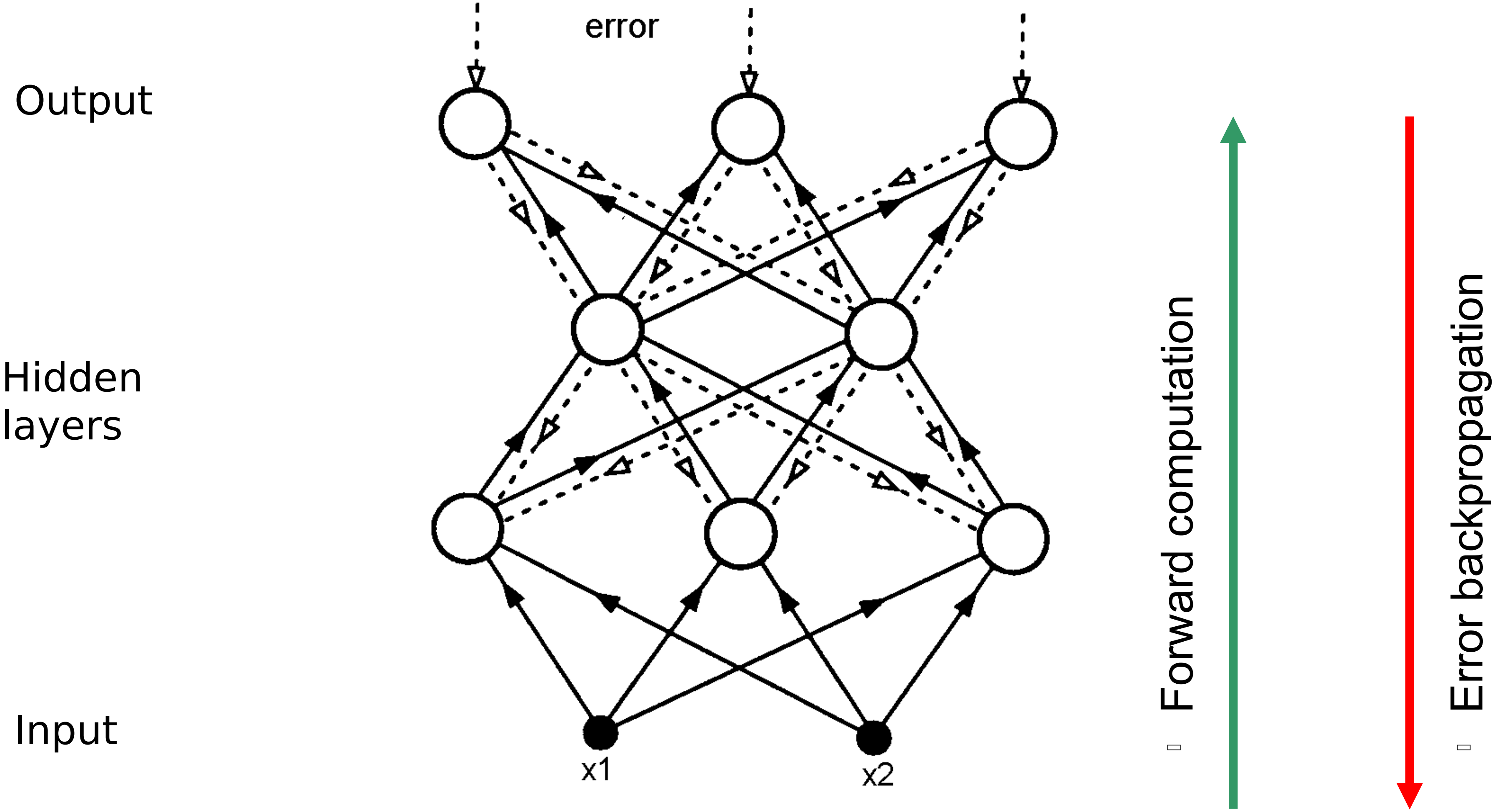
Input



Forward computation

Error backpropagation

A schematic view of the error backprogation algorithm for a 3-layers neural network



Output

Hidden layers

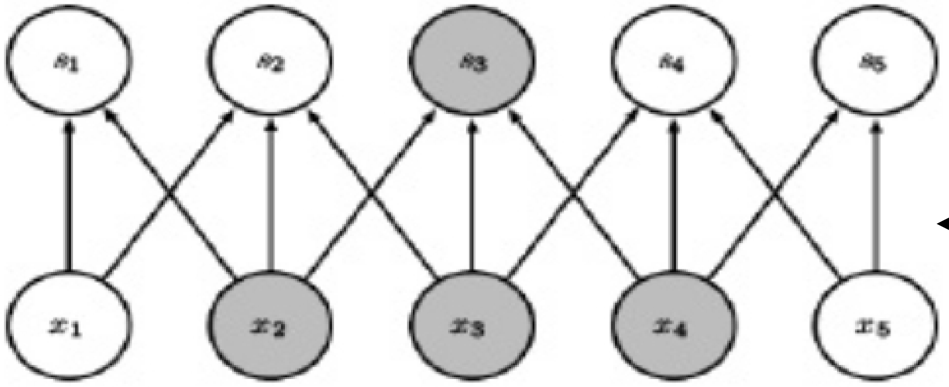
Input

Forward computation

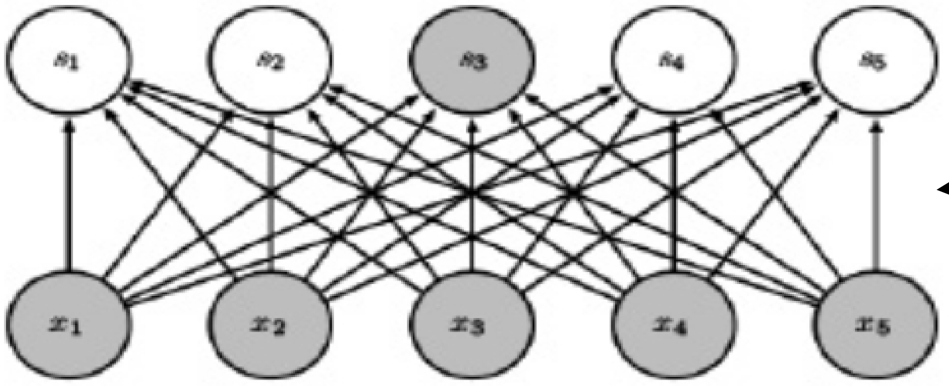
Error backpropagation

<u>But learning is problematic with deep fully connected neural networks</u> ...

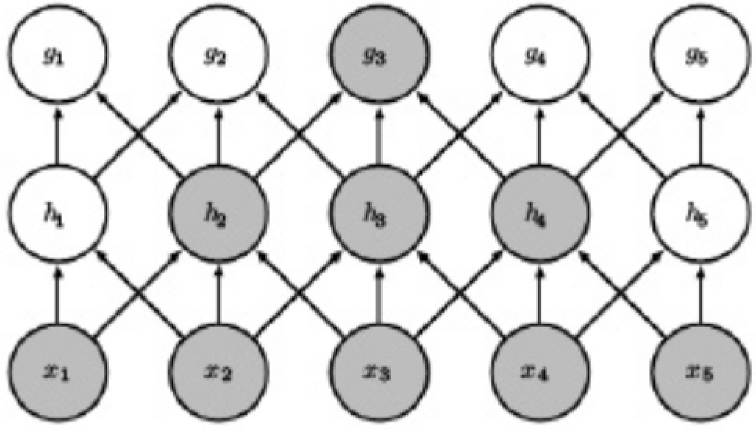
Convolutional neural networks and “smart” learning make feasible DNN training - 1



Convolutional network: sparsity of connections



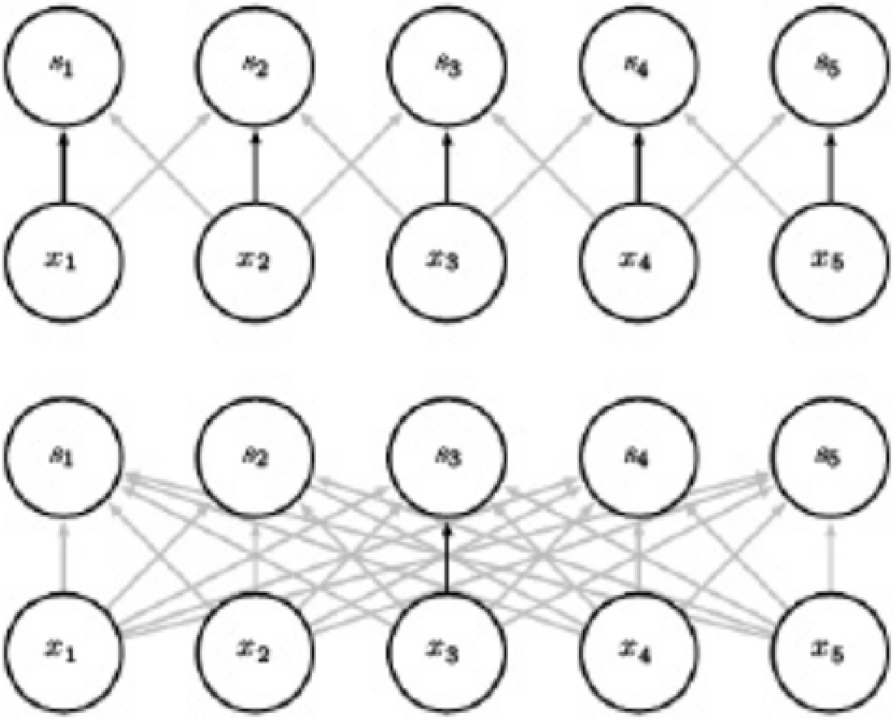
Fully connected network



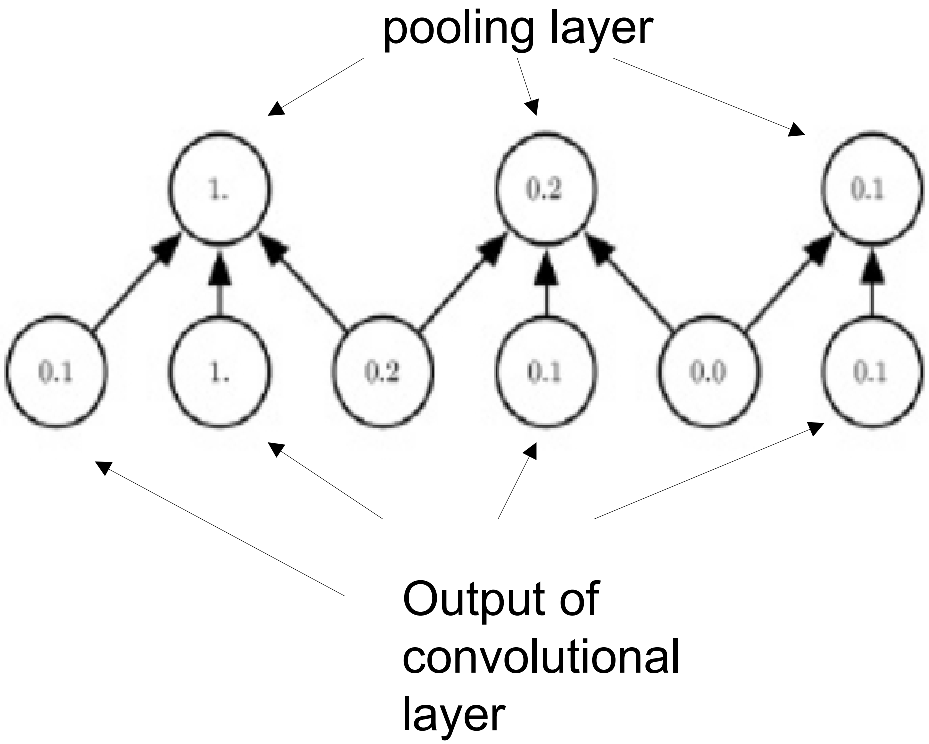
Locality does not mean reducing the receptive field ...

Convolutional neural networks and "smart" learning make feasible DNN training - 2

Convolution allows parameter sharing

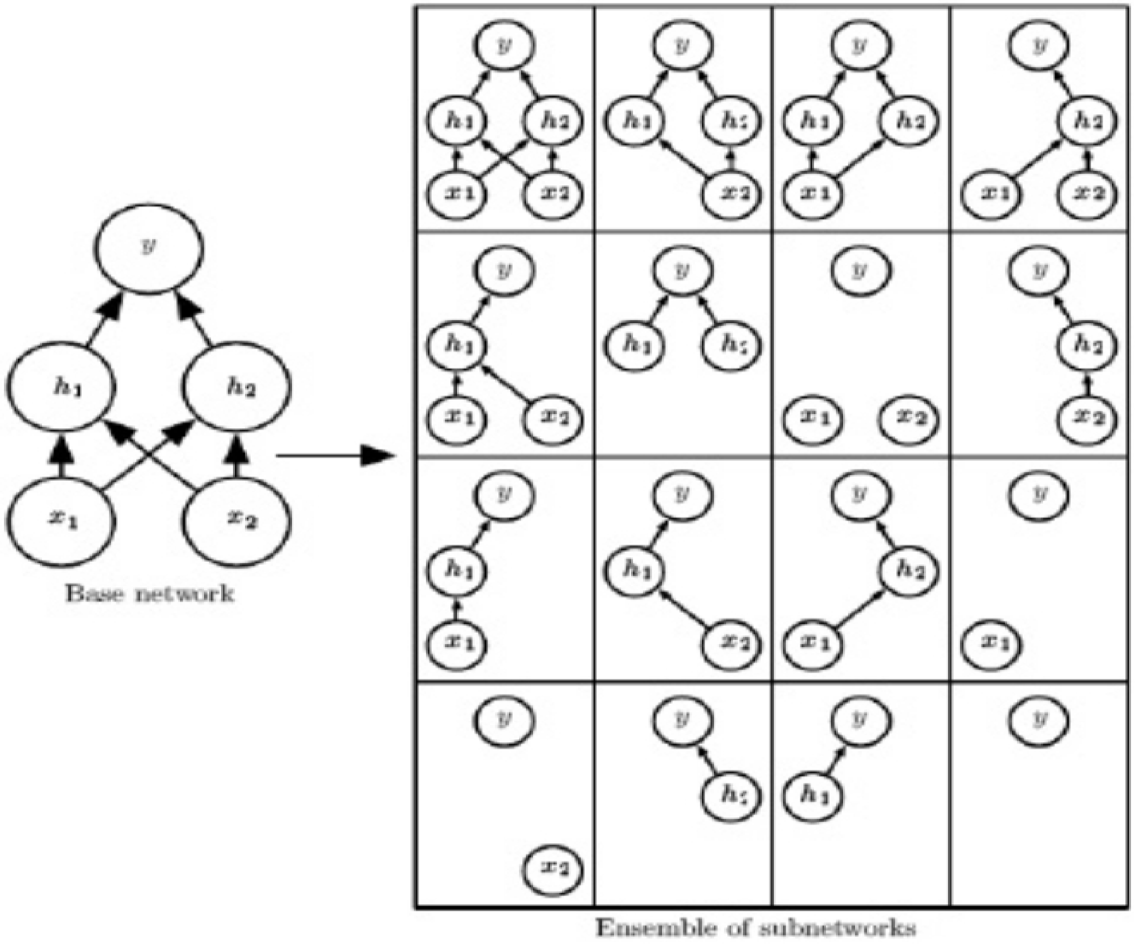


Pooling allows invariant transformations and *generalization*


pooling layer

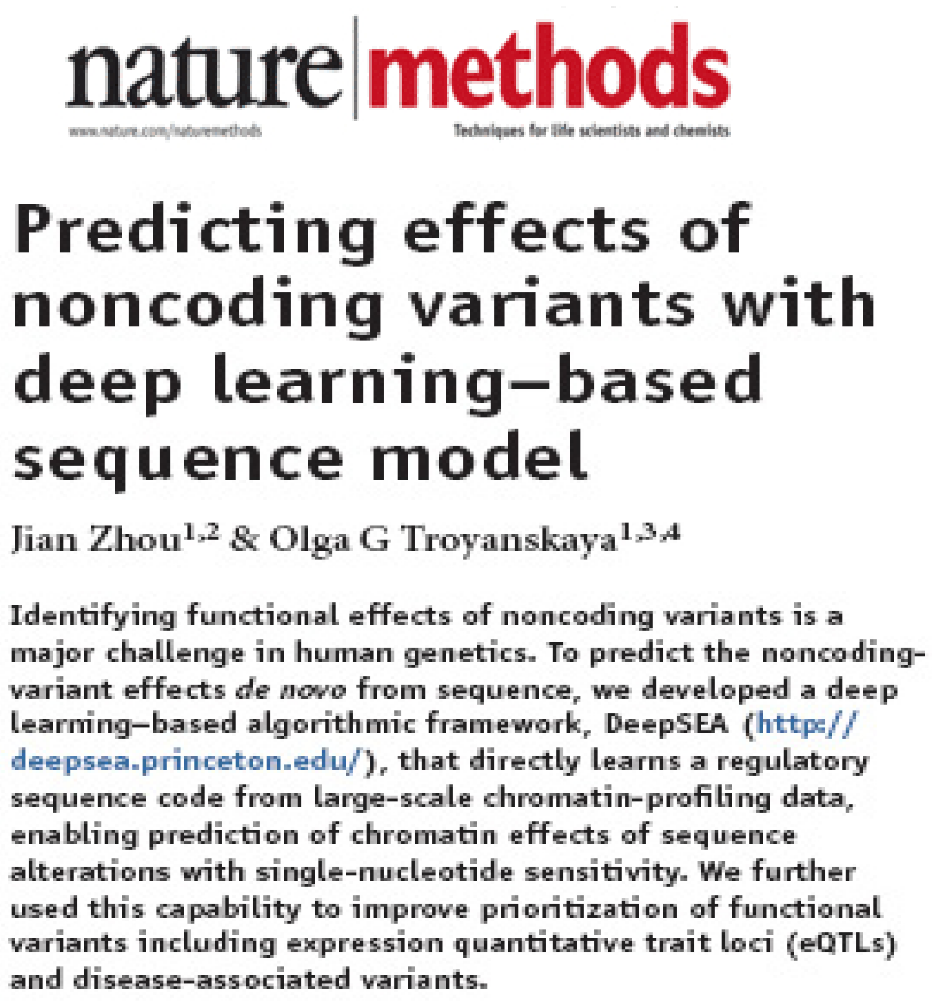
Output of convolutional layer

Convolutional neural networks and “smart” learning make feasible DNN training - 3



Stochastic gradient descent and *dropout learning algorithms* (Srivastava et al., 2014) allow fast training of big deep networks.

The DeepSea method for interpreting non coding variants



nature methods

www.nature.com/naturemethods

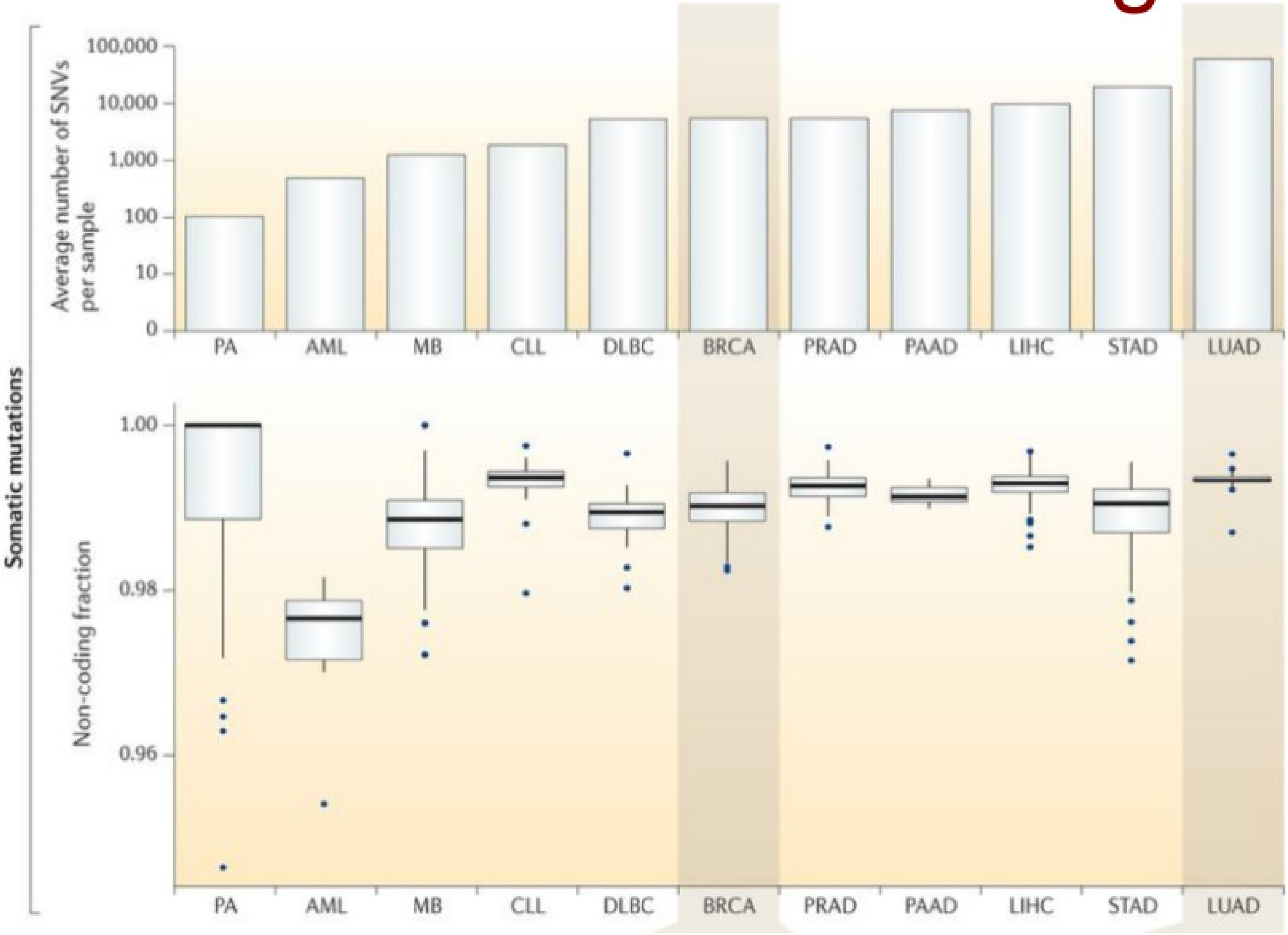
Techniques for life scientists and chemists

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

Jian Zhou[1,2] & Olga G Troyanskaya[1,3,4]

Identifying functional effects of noncoding variants is a major challenge in human genetics. To predict the noncoding-variant effects *de novo* from sequence, we developed a deep learning–based algorithmic framework, DeepSEA (http://deepsea.princeton.edu/), that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity. We further used this capability to improve prioritization of functional variants including expression quantitative trait loci (eQTLs) and disease-associated variants.

Almost all single nucleotide variants in cancer are noncoding



Khurana *Nature Reviews Genetics* 2016

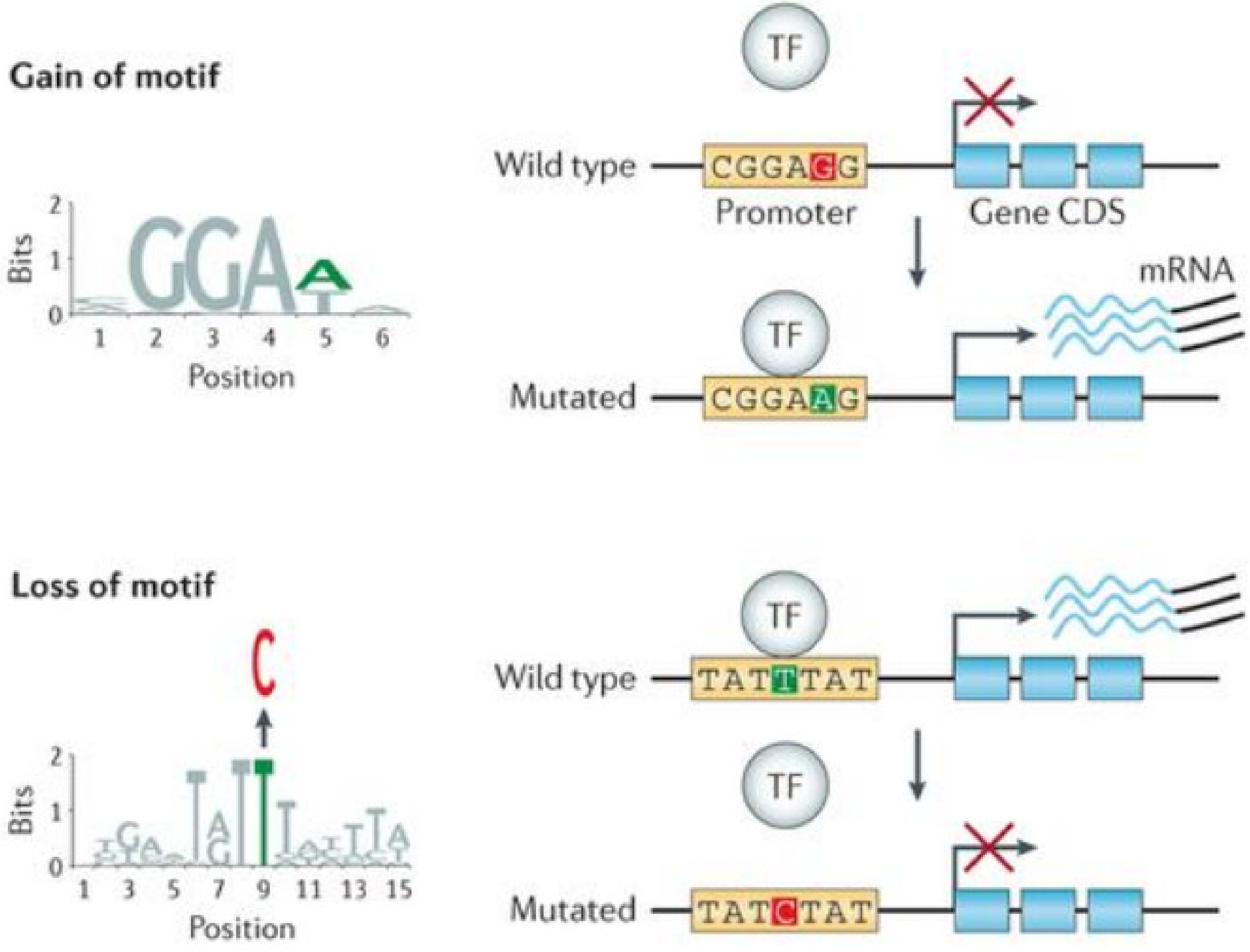
However, very few of these are driver mutations

Ways a noncoding variant can be functional

- Disrupt DNA sequence motifs
  - Promoters, enhancers
- Disrupt miRNA binding
- Mutations in introns affect splicing
- Indirect effects from the above changes

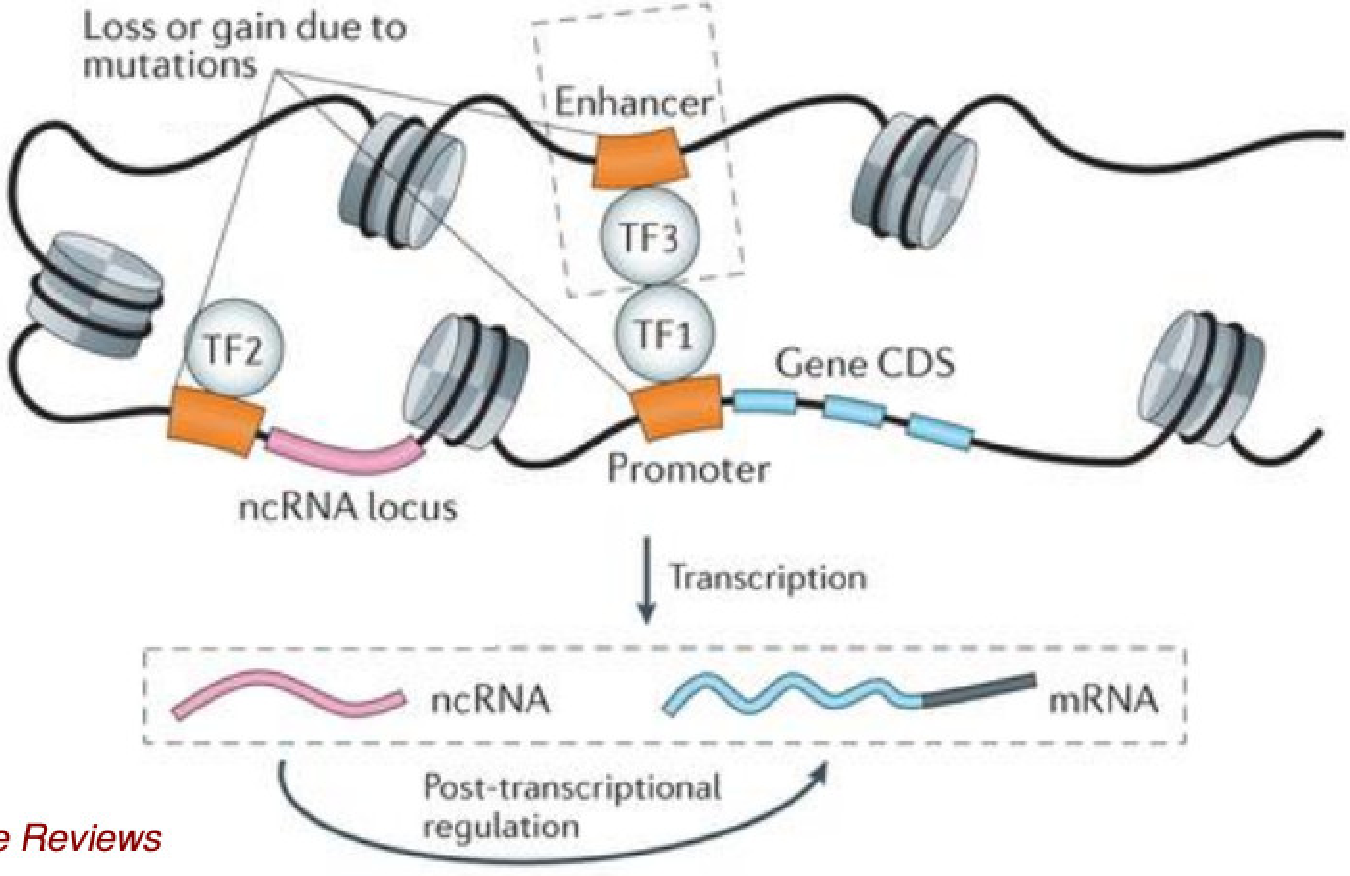
Examples in Ward and Kellis *Nature Biotechnology* 2012

Variants altering motifs



Khurana *Nature Reviews Genetics* 2016

Variants affect proximal and distal regulators



Khurana *Nature Reviews Genetics* 2016

DeepSEA

- Given:
  - A sequence variant and surrounding sequence context

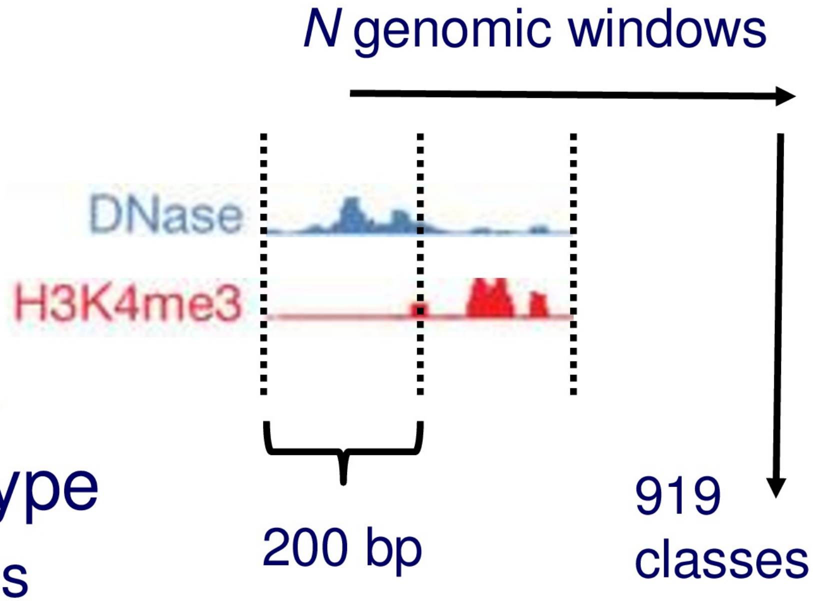
- Do:
  - Predict TF binding, DNase hypersensitivity, and histone modifications in multiple cell and tissue types
  - Predict variant functionality

Cell variables to be predicted:

- DNase hypersensitivity
- TF binding
- Histone modification

Classifier input and output

- Output
  - 200 bp windows of genome
  - Label 1 if window contains peak
  - Label for each epigenetic data type
    - Multiple types of epigenetic features
    - Multiple types of cells and tissues



Roadmap Epigenomics Consortium *Nature* 2015

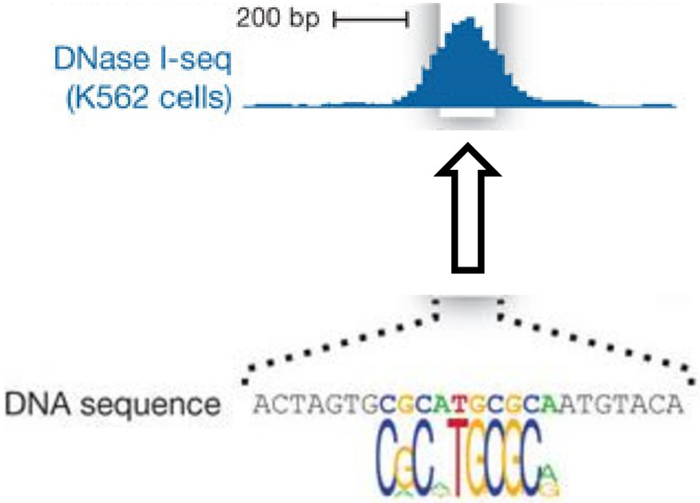
- Input: 1000 bp DNA sequence centered at window

$$x_i = \begin{array}{lcccccccc} \texttt{index} & \texttt{1} & \dots & \texttt{401} & \texttt{402} & \texttt{403} & \dots & \texttt{1000} \\ \texttt{A} & \texttt{0} & & \texttt{1} & \texttt{0} & \texttt{0} & & \texttt{0} \\ \texttt{C} & \texttt{0} & & \texttt{0} & \texttt{0} & \texttt{0} & & \texttt{1} \\ \texttt{G} & \texttt{1} & & \texttt{0} & \texttt{1} & \texttt{1} & & \texttt{0} \\ \texttt{T} & \texttt{0} & & \texttt{0} & \texttt{0} & \texttt{0} & & \texttt{0} \end{array}$$

Desired properties for epigenomic classifier

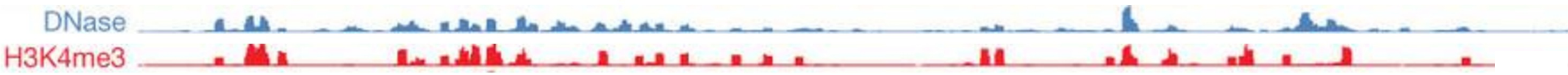
- Learn preferences of DNA-binding proteins
  - Locally: “motifs” and other simple sequence patterns
  - Sequence context: “*cis*-regulatory modules”

- Support nonlinear decision boundaries



Neph *Nature* 2012

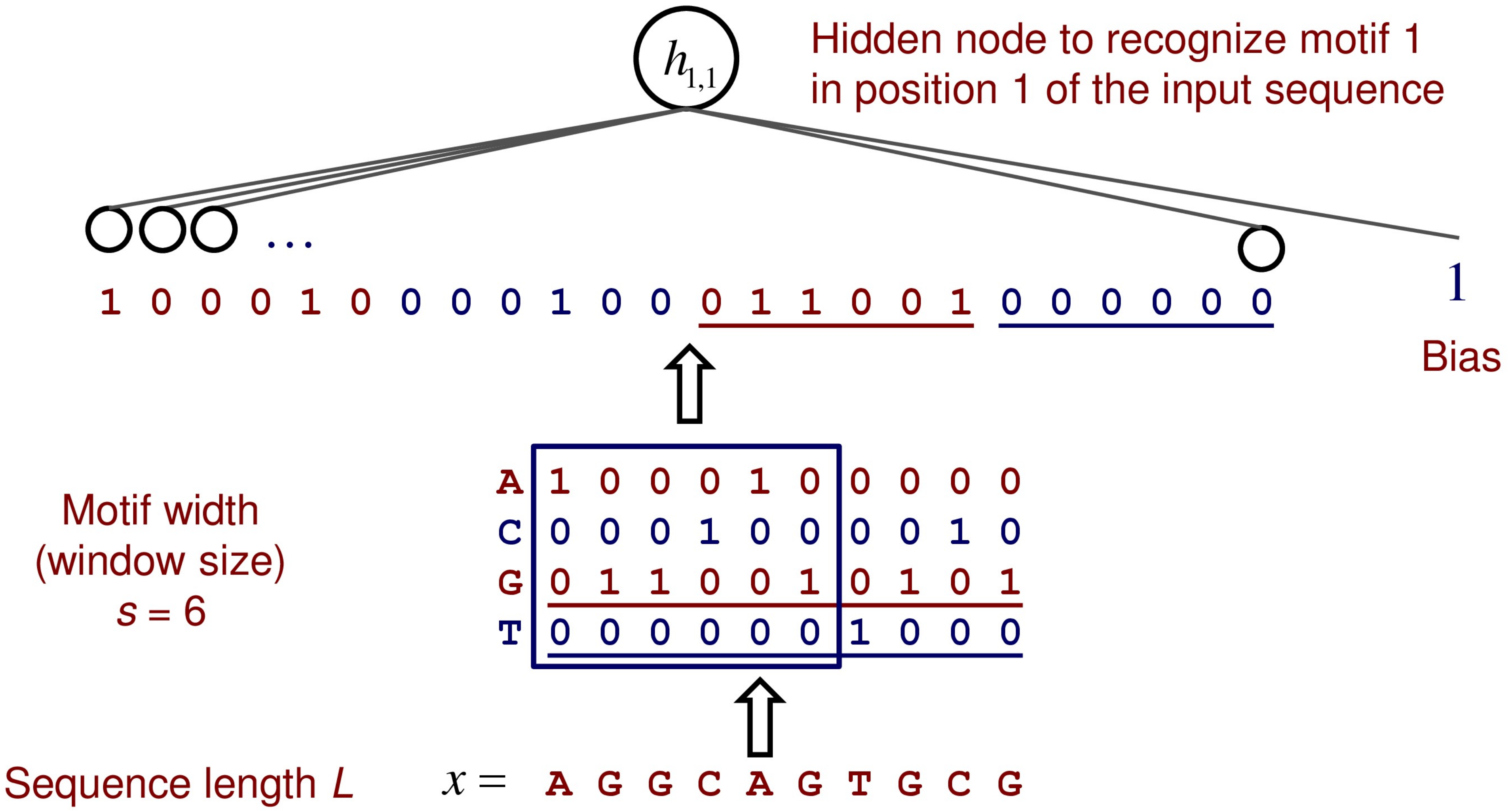
- Multiple, related prediction tasks



Roadmap Epigenomics Consortium *Nature* 2015

First hidden layer

- First hidden layer scans input sequence
- Activation function fires if “motif” is recognized



Hidden node to recognize motif 1 in position 1 of the input sequence

1 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0

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

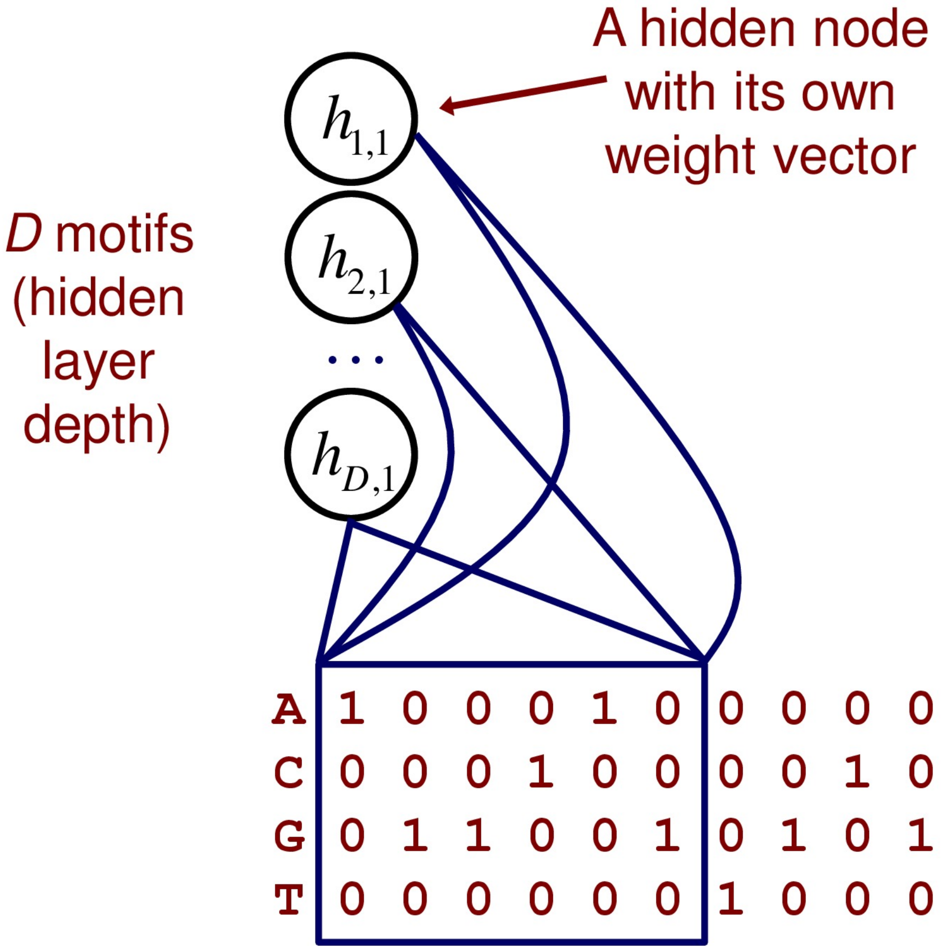
Motif width
(window size)
s = 6

Sequence length *L*

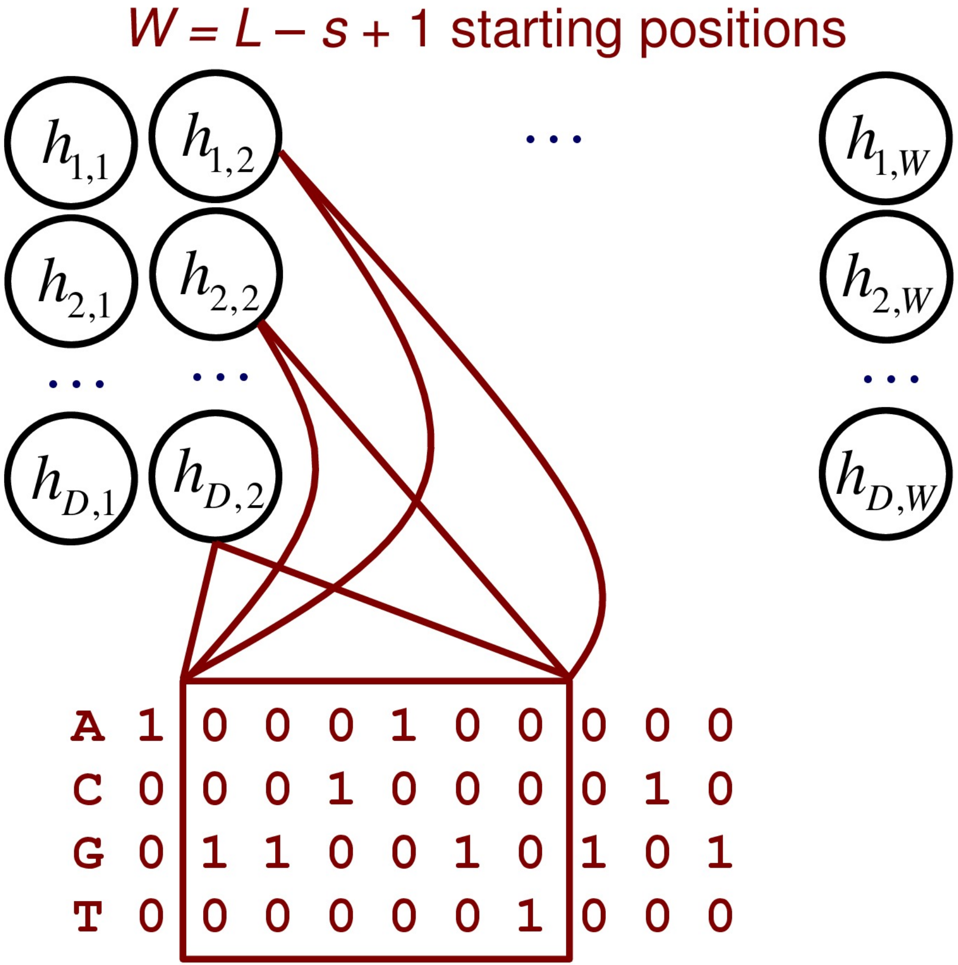
$x =$ A G G C A G T G C G

First hidden layer

- Multiple hidden nodes to recognize different motifs at a particular position
- Check for motif at each position in sequence


A hidden node with its own weight vector
D motifs (hidden layer depth)
$h_{1,1}$
$h_{2,1}$
$\cdots$
$h_{D,1}$

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



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

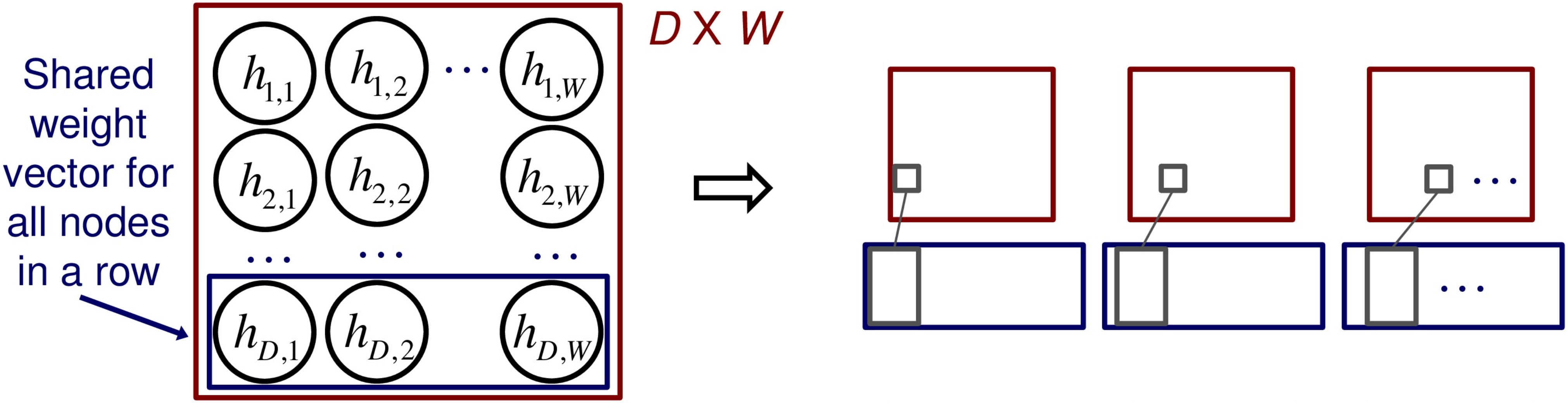
First layer problems

- We already have a *lot* of parameters
  - Each hidden node has its own weight vector

- We're attempting to learn different motifs at each starting position

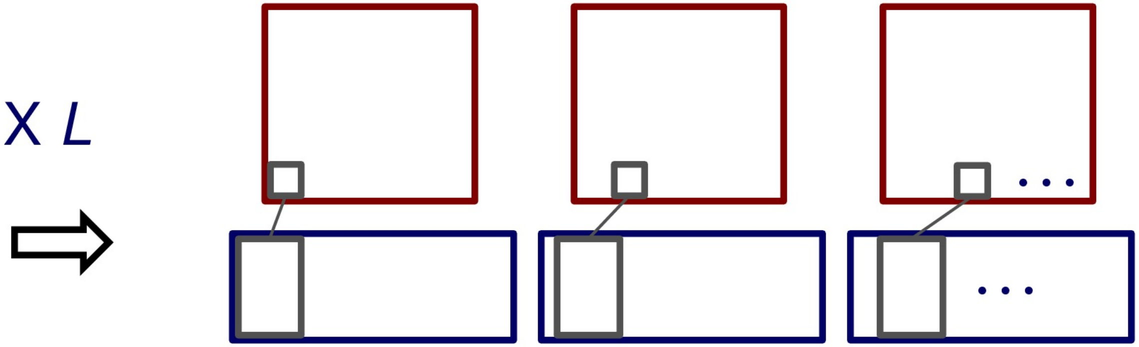
Convolutional layers

- Input sequence and hidden layer as matrices
- Share parameters for all hidden nodes in a row
  - Search for same motif at different starting positions



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

4 X *L*



Pooling layers

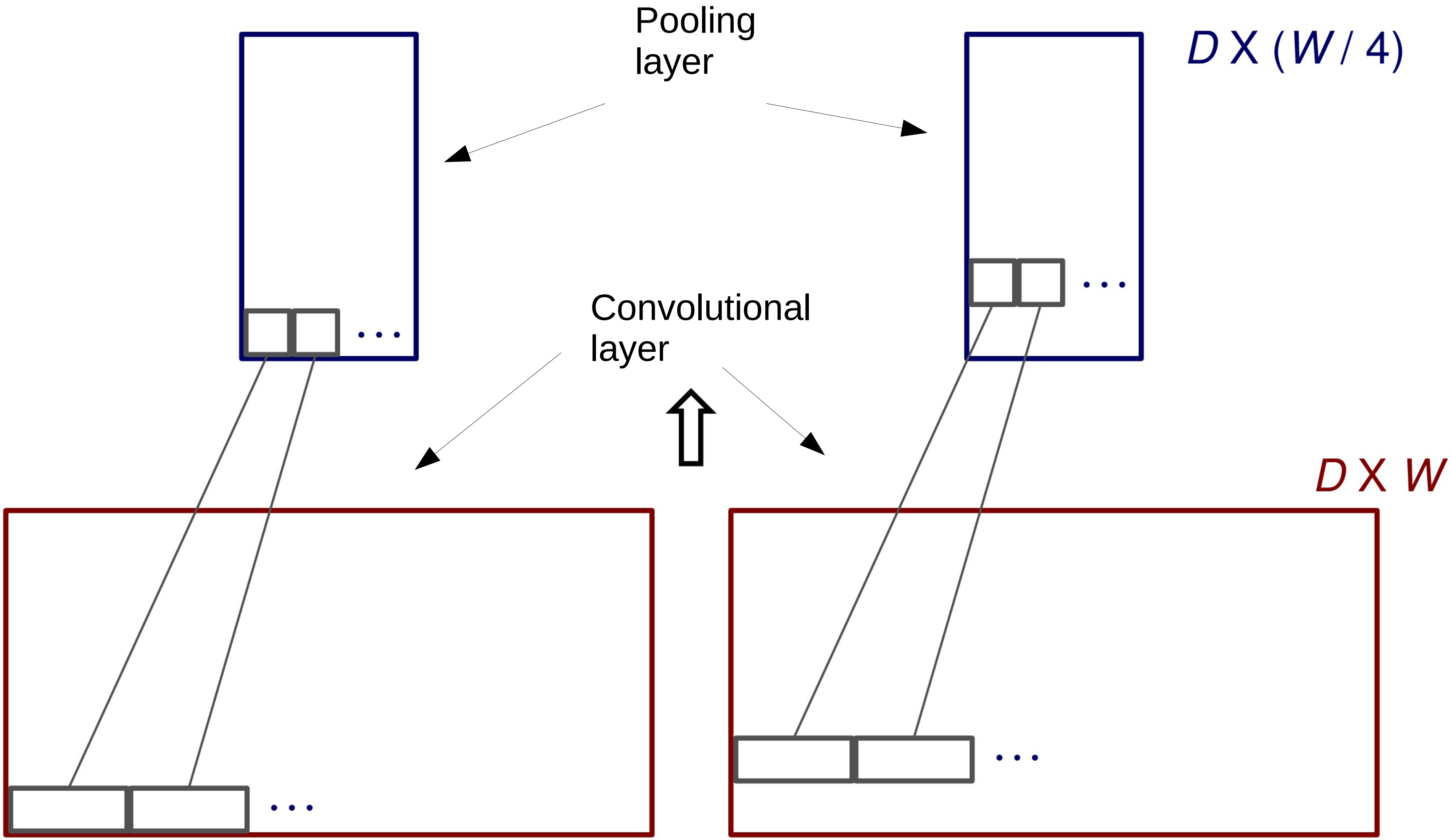
- Account for sequence context

- Multiple motif matches in a *cis*-regulatory module

- Search for patterns at a higher spatial scale
  - Fire if motif detected anywhere within a window

Pooling layers

- Take max over window of 4 hidden nodes



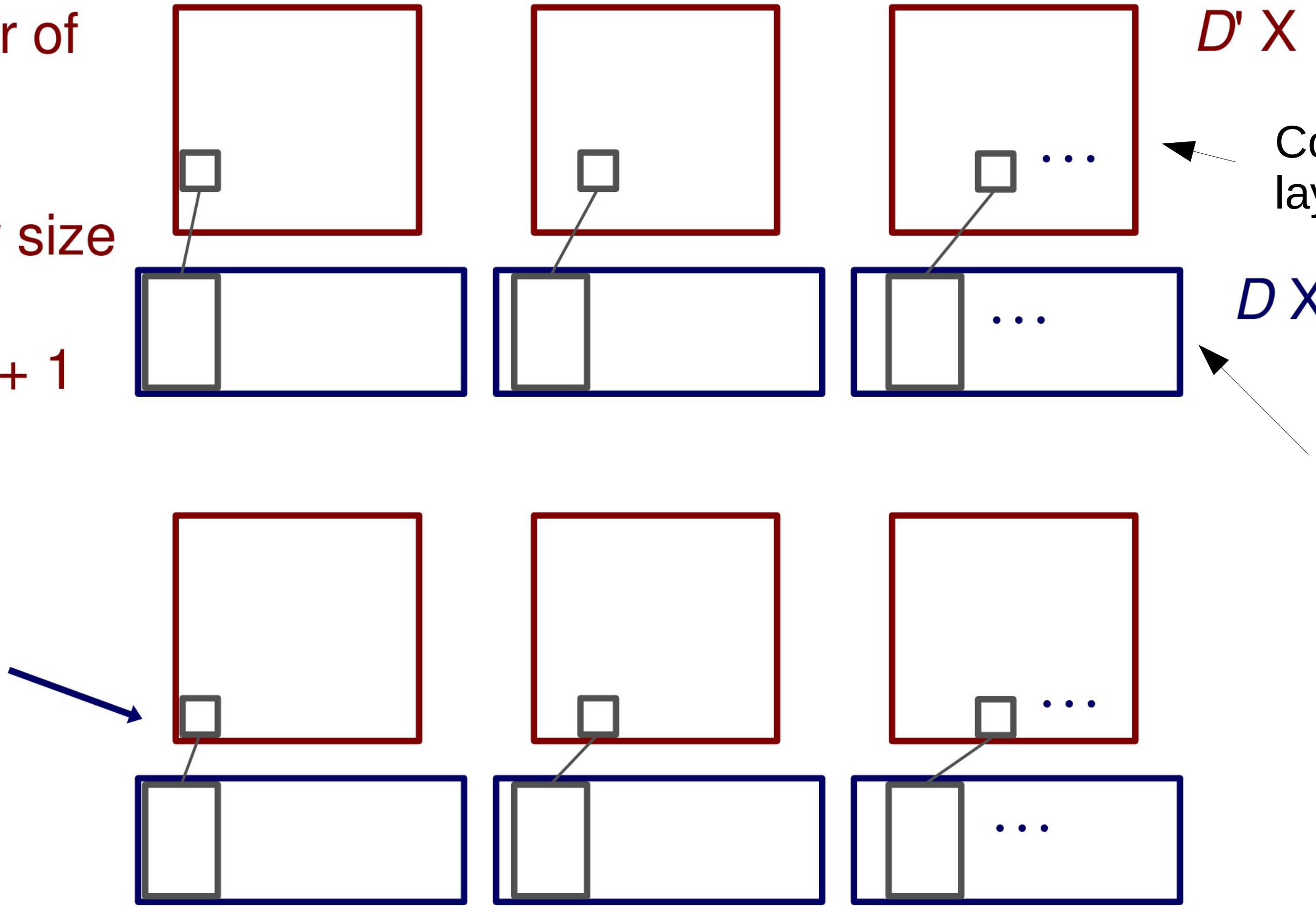
Subsequent hidden layers

- Next convolutional hidden layer on top of pooling layer

D' is new number of patterns

s' is new window size

$W' = (W / 4) - s' + 1$



D' X *W* '

Convolutional layer

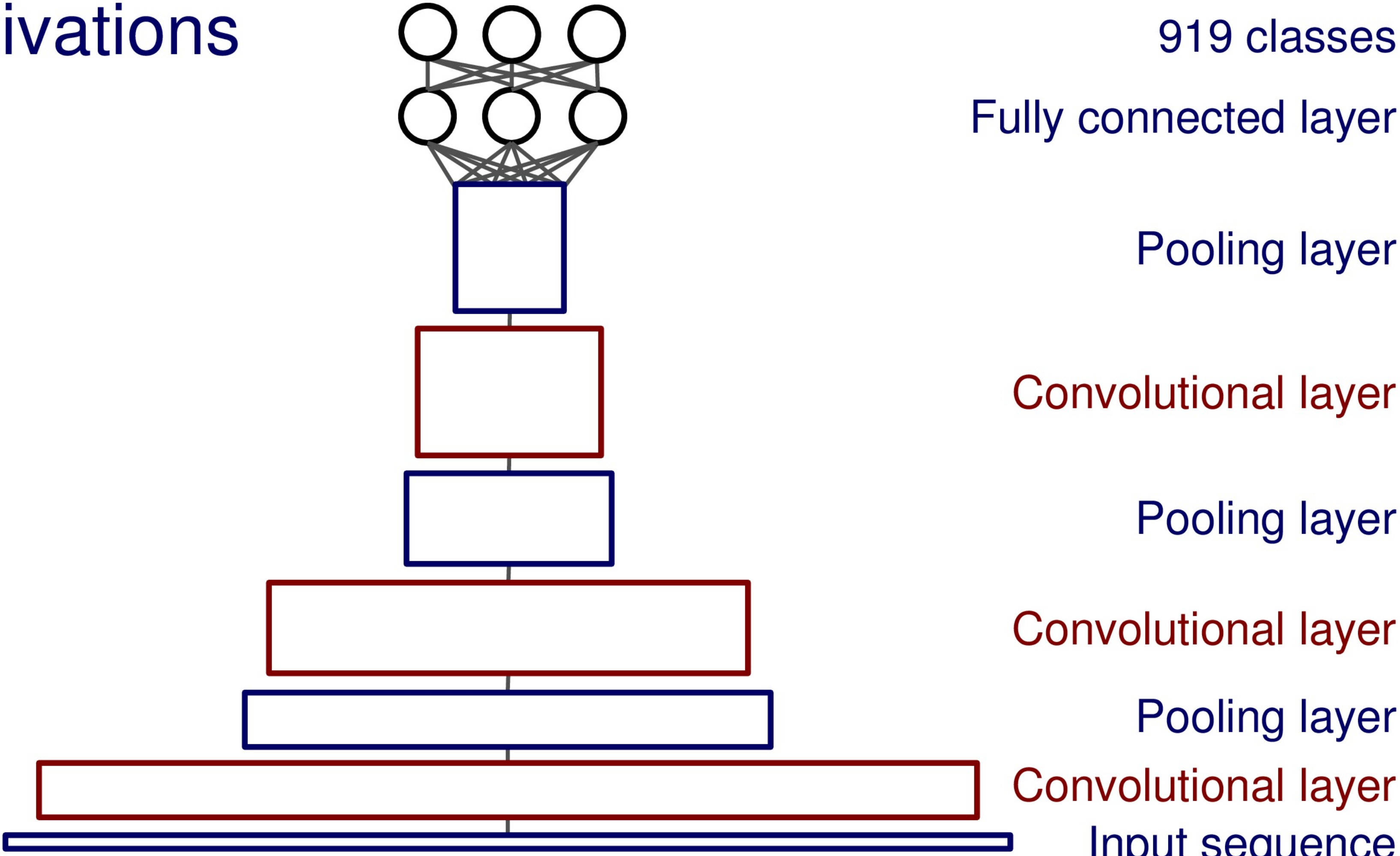
D X (W / 4)

Pooling layer

Once again, shared weight vector for all nodes in a row

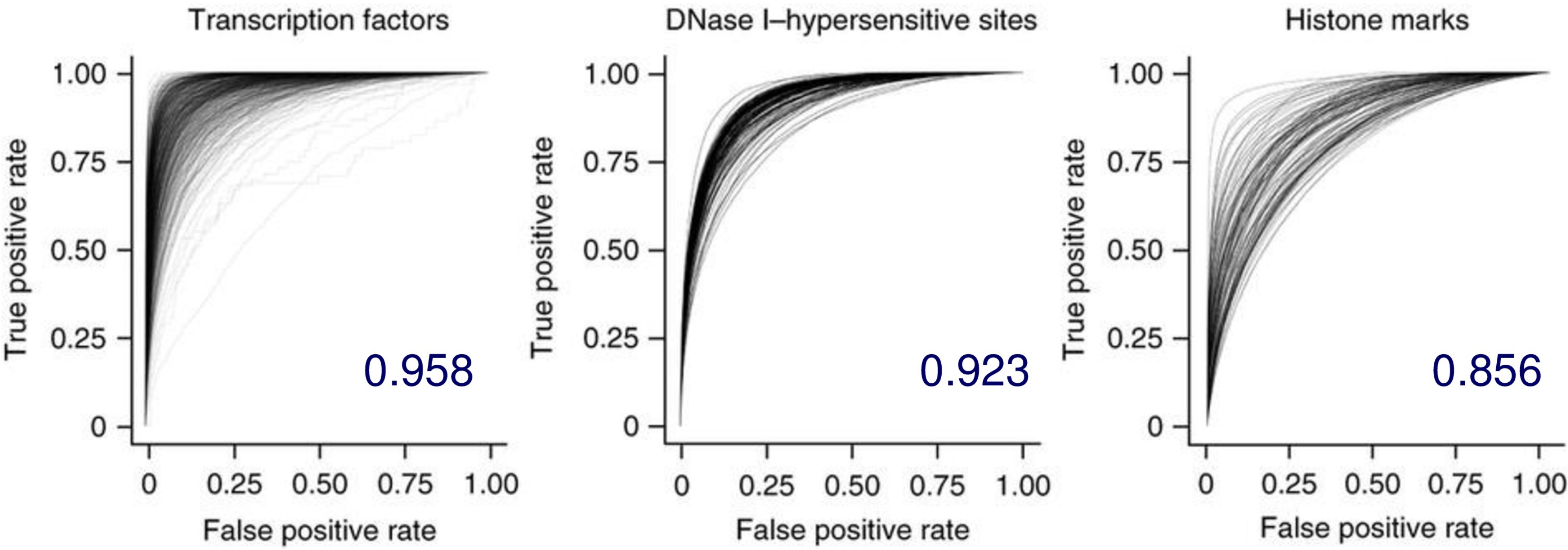
Full DeepSEA neural network

- Multitask output makes simultaneous prediction for each type of epigenetic data
- ReLU activations



Predicting epigenetic annotations

- Compute median AUC ROC for three types of classes



Zhou and Troyanskaya *Nature Methods* 2015

Predicting functional variants

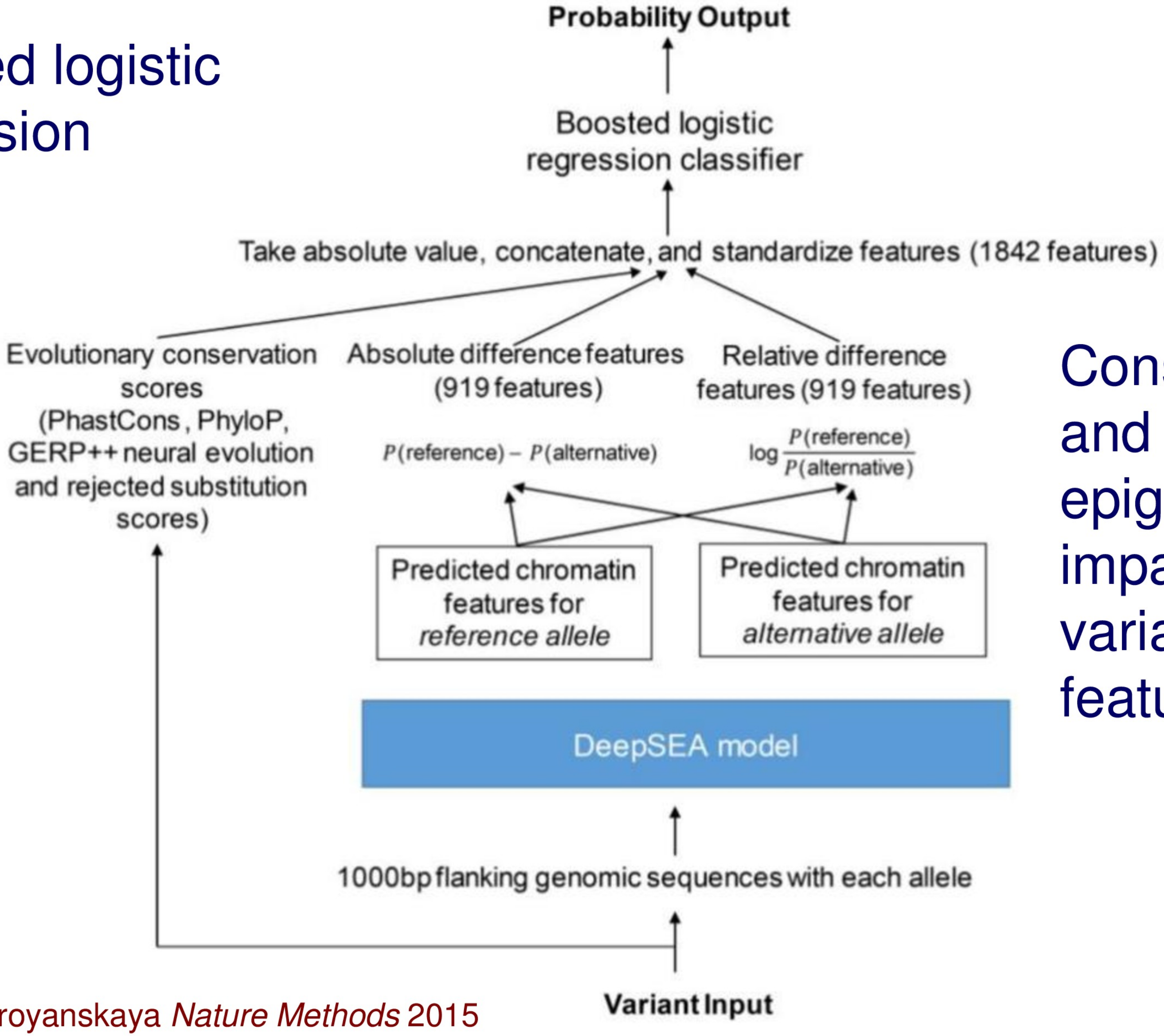
- Can predict epigenetic signal for any novel variant (SNP, insertion, deletion)

- Define novel features to classify variant functionality
  - Difference in probability of signal for reference and alternative allele

- Train on SNPs annotated as regulatory variants in GWAS and eQTL databases

Predicting functional variants

Boosted logistic regression



Probability Output

Boosted logistic regression classifier

Take absolute value, concatenate, and standardize features (1842 features)

Evolutionary conservation scores (PhastCons, PhyloP, GERP++ neural evolution and rejected substitution scores)

Absolute difference features (919 features)

$P(\text{reference}) - P(\text{alternative})$

Relative difference features (919 features)

$\log \frac{P(\text{reference})}{P(\text{alternative})}$

Predicted chromatin features for *reference allele*

Predicted chromatin features for *alternative allele*

DeepSEA model

1000bp flanking genomic sequences with each allele

Variant Input

Conservation and predicted epigenetic impact of variant as features

Zhou and Troyanskaya *Nature Methods* 2015

DeepSEA summary

- Ability to predict how unseen variants affect regulatory elements
- Accounts for sequence context of motif
- Parameter sharing with convolutional layers
- Multitask learning to improve hidden layer representations

- Does not extend to new types of cells and tissues
- AUC ROC is misleading for evaluating genome-wide epigenetic predictions

State-of-the-art ML methods for the prediction of deleterious/pathogenic variants

- CADD (Kircher, et al. 2014)
- GWAVA (Ritchie et al 2014)
- DeepSEA (Zhou & Troyanskaya, 2015)
- FATHMM-MKL (Shibab et al. 2015)
- Eigen (Ionita-Laza et al. 2016)
- LINSIGHT (Huang et al. 2017)

Quite surprisingly none of the above methods (apart from GWAVA) use imbalance-aware learning strategies

References - 1

- M. A. Rubin, ‘‘Make precision medicine work for cancer care,’’ *Nature*, vol. 520, no. 7547, pp. 290–291, 2015.
- M. Leung, A. Delong, B. Alipanahi, B. Frey, Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets *Proceedings of the IEEE* 104(1), 2016
- H. Y. Xiong et al., The human splicing code reveals new insights into the genetic determinants of disease, *Science*, vol. 347, no. 6218, 2014
- P. M. Visscheret al. 10 Years of GWAS Discovery: Biology, Function, and Translation, *Amer. J. Human Genetics*, Vol. 101, Issue 1, 2017
- J. Kruppa, A. Ziegler, and I. R. Konig, Risk estimation and risk prediction using machine-learning methods, *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012.
- M. Cooper et al., ‘‘Distribution and intensity of constraint in mammalian genomic sequence,’’ *Genome Res*., vol. 15, no. 7, pp. 901–913, 2005.
- K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, ‘‘Detection of nonneutral substitution rates on mammalian phylogenies,’’ *Genome Res*., vol. 20, no. 1, pp. 110–121, 2010.
- D. G. MacArthur et al., ‘‘Guidelines for investigating causality of sequence variants in human disease,’’ *Nature*, vol. 508, no. 7497, pp. 469–476, 2014.
- D. Quang, Y. Chen, and X. Xie, ‘‘DANN: A deep learning approach for annotating the pathogenicity of genetic variants,’’ *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2014.
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46, 310–315, 2014
- M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, ‘‘Deep learning of the tissue-regulated splicing code,’’ *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014.
- B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, ‘‘Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,’’ *Nature Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015
- J. Zhou, O.G. Troyanskaya Predicting effects of noncoding variants with deep learning-based sequence model *Nat. Methods*, 12 pp. 931-934, 2015
- K. Chaudhary et al. “Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer.” *Clinical Cancer Research* 24 (6), 2018

References - 2

- A. Burga and B. Lehner, ‘‘Beyond genotype to phenotype: Why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience,’’ *FEBS* J., vol. 279, no. 20, pp. 3765–3775, 2012.
- J. Shim et al., ‘‘Nanopore-based assay for detection of methylation in double-stranded DNA fragments,’’ *ACS Nano*, vol. 9, no. 1, pp. 290–300, 2015.
- B. Treutlein, et al. ‘‘Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing,’’ *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 13, pp. E1291–E1299, 2014.
- O. Stegle, S. A. Teichmann, and J. C. Marioni, ‘‘Computational and analytical challenges in single-cell transcriptomics,’’ *Nature Rev. Genetics*, vol. 16, no. 1, pp. 133–145, 2015.
- A. Kreimer et al. Meta-analysis of massive parallel reporter assay enables functional regulatory elements prediction, *bioRxiv* doi.10.1101/202002, 2017
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. *Nat Methods* 7(4):248-249 (2010)
- Bendl, J., Musil, M., Stourac, J., Zendulka, J., Damborsky, J., Brezovsky, J., 2016: PredictSNP2: A unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLOS Computational Biology* 12: e1004962, 2016
- LeCun, Y., Bengio, Y. , Hinton, G. Deep learning *Nature* 521, 2015
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 , 1929–1958, 2014.
- Ching T et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15: 20170387, 2018.
- Khurana E. et al. Role of non-coding sequence variants in cancer *Nature Reviews Genetics* 17, 2016
- Roadmap Epigenomic Consortium, Integrative analysis of 111 reference human epigenomes *Nature* 518, 2015
- D. Cox, R. Platt, and F. Zhang, Therapeutic genome editing: Prospects and challenges, *Nature Med.*, vol. 21, no. 2, pp. 121–131, 2015

References - 3

- Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nature Methods* 11, 294–6, 2014.
- Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics* 49, 618–24, 2017
- Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics* 48, 214–20, 2016
- Shihab, H. A. et al. An integrative approach to predicting the functional effects of noncoding and coding sequence variation. *Bioinformatics* 31, 1536–43, 2015
- Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics* 49, 618–24, 2017
- Smedley, D. et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics* 99, 595–606, 2016
- Schubach, M., Re, M., Robinson, P. N. & Valentini, G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Scientific Reports* 7, 2959, 2017.

Acknowledgments

Many thanks to:
Martin Kircher (Berlin Institute of Health) for providing the slides about the CADD method
and to:
Anthony Gitter (University of Wisconsin-Madison, USA) for providing the slides about DeepSea.

But above all:

Thank you for your attention !

Computer Science
Department

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Anacleto
Lab

Computational Biology and Bioinformatics