UNIVERSITY OF EASTERN FINLAND

# Deep neural network configurations for DNA motif analysis

Juha Mehtonen, MSc, PhD student

Institute of Biomedicine, School of Medicine, University of Eastern Finland

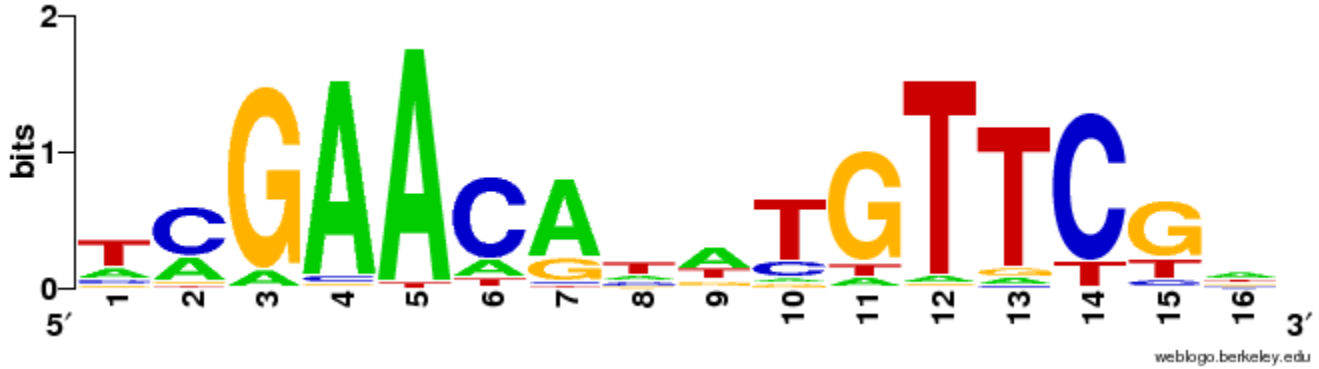
# Presentation overview

- *De novo* motif discovery
- DeepBind
- DeepSea & DanQ


# *De novo* motif discovery

# *De novo* motif discovery

- Given multiple input sequences, attempt to identify one or more candidate motifs (recurring patterns)
- Input: DNA sequences of varying lengths
  - ~20 for identifying sequence motifs
  - ~200-1000 for identifying sequence functions (deep learning applications)
- Input vocabulary consists of letters A, C, G and T.



# *De novo* motif discovery

- **Position weight matrix** (PWM)
  - Commonly used representation of motifs in biological sequences.
  - Can be used to score whether a sequence matches the motif of interest.
- Simple way to create a PWM:
  1. Given a set of aligned sequences, calculate position frequency matrix **X** where each element $\mathbf{x}_{i,j}$ corresponds to the frequency of nucleotide (or amino acid) *i* at position *j*.
  2. Compute a position probability matrix by dividing **X** with the number of input sequences.
  3. Finally, create the PWM by dividing each row (nucleotide probabilities) by a background probability *b* and calculating logarithm of the matrix.
  4. Resulting PWM represents log likelihoods of nucleotides appearing at specific positions.

# *De novo* motif discovery

- PWM model doesn't take into account possible variable spacing or gaps in the motif.
- Appearance of a nucleotide at one position of the sequence does not depend on the nucleotides that appear on other positions of the site.
- More complex models needed
  - Markov chain, Bayes, deep learning…

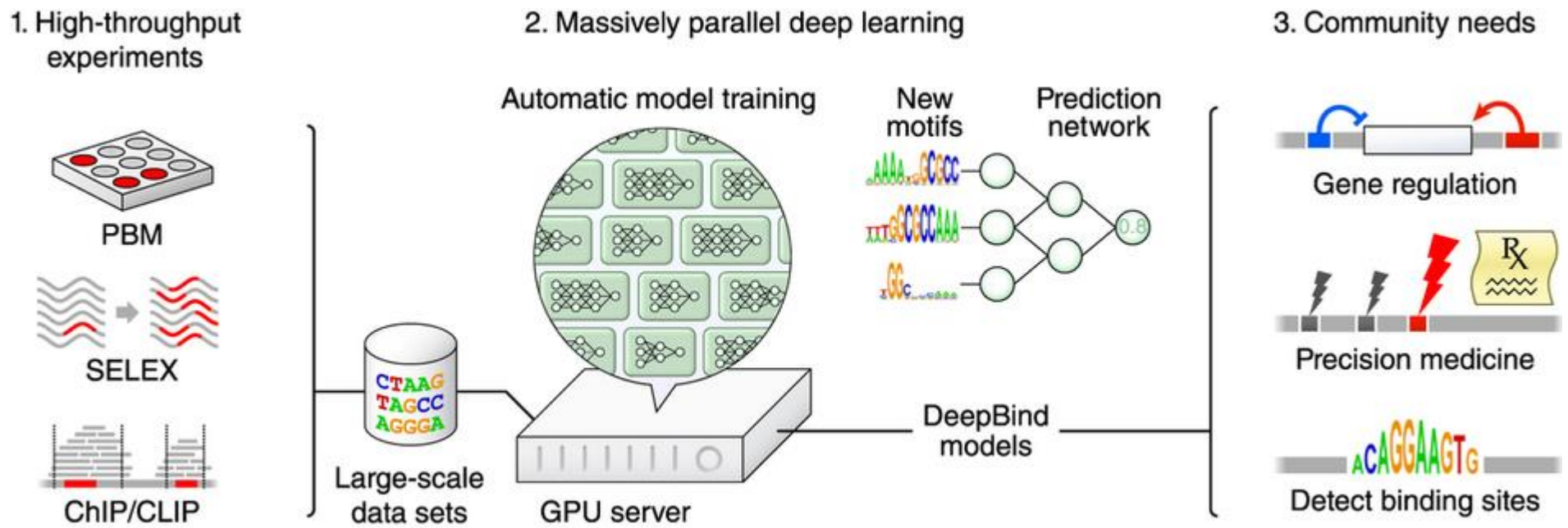

# DeepBind

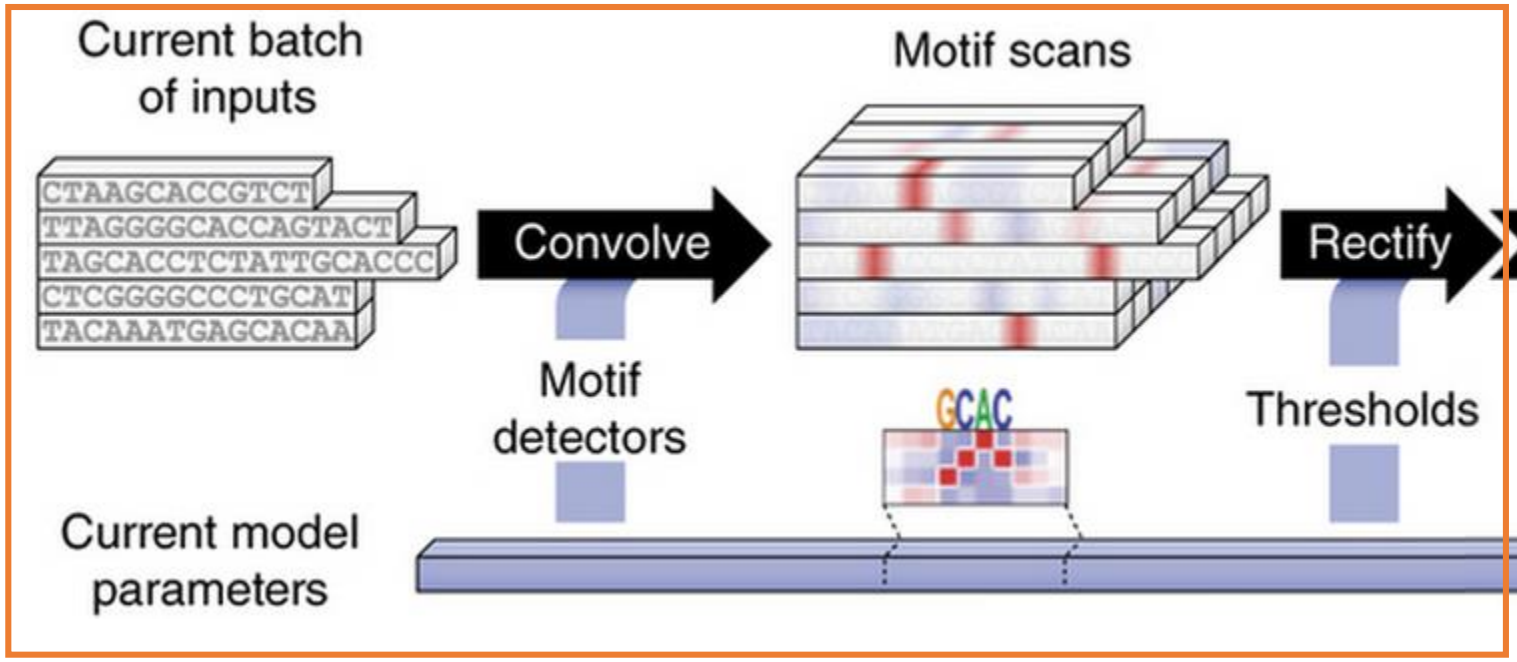
# DeepBind

- Learns DNA binding pattern (motif) from DNA sequences.
- The model can then be used to predict
  - if new unknown sequences have the same binding pattern
  - how variations in DNA sequence affect binding with a specific sequence
- One model learns one motif, multiple models needed to learn multiple motifs.
- Alipanahi *et al*. 2015, Nature Biotechnology

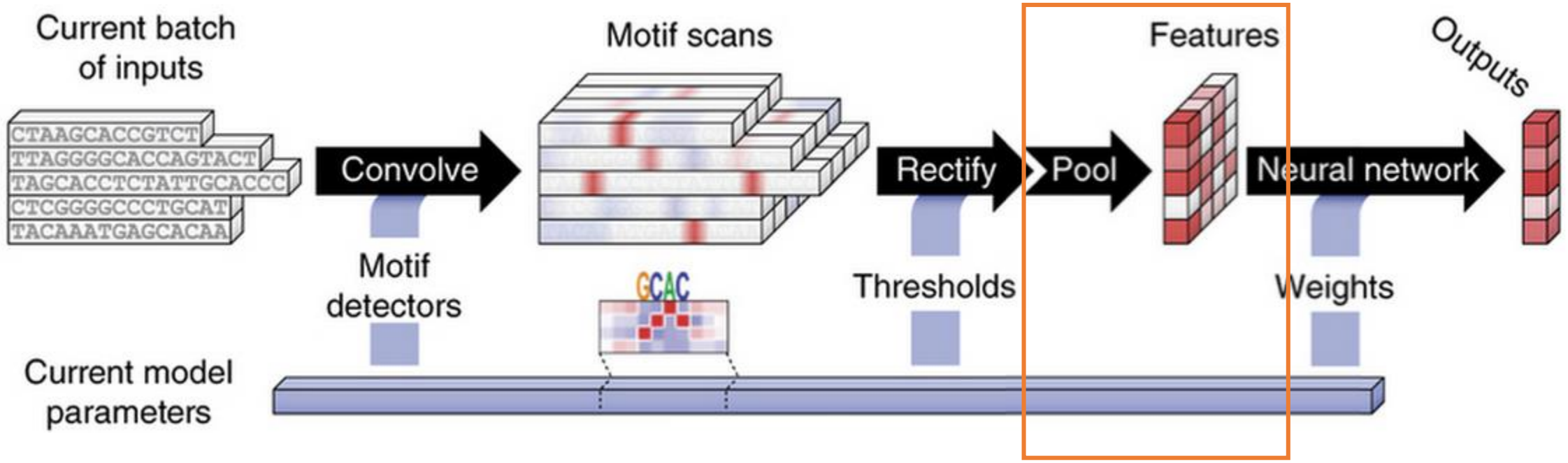
# DeepBind

- Deep convolutional neural network for discovering patterns from sequences.

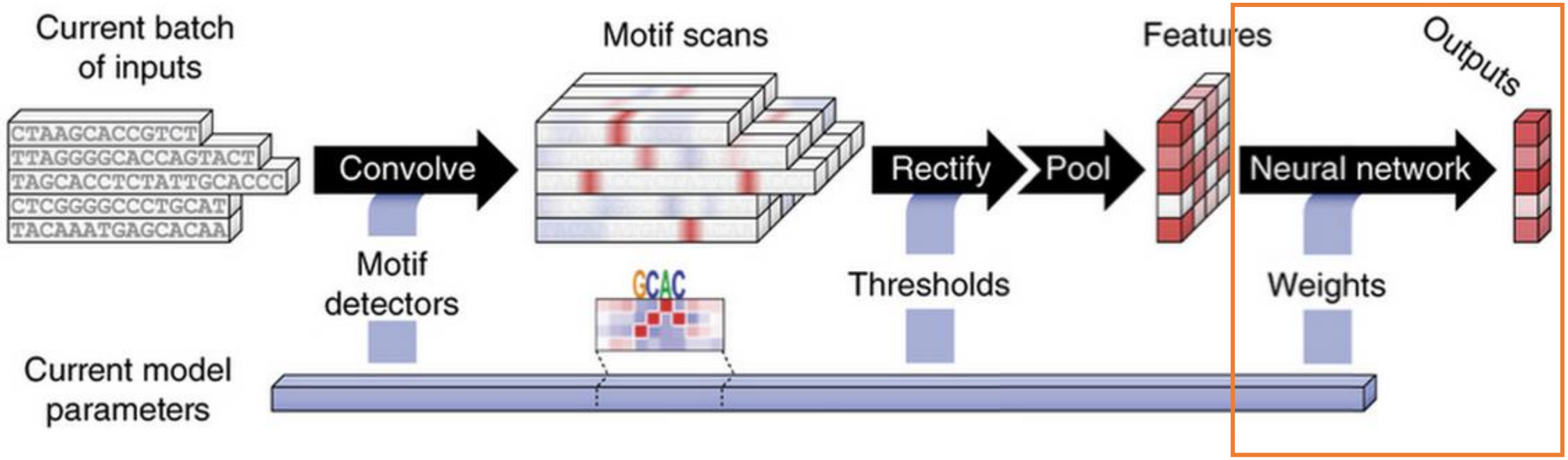




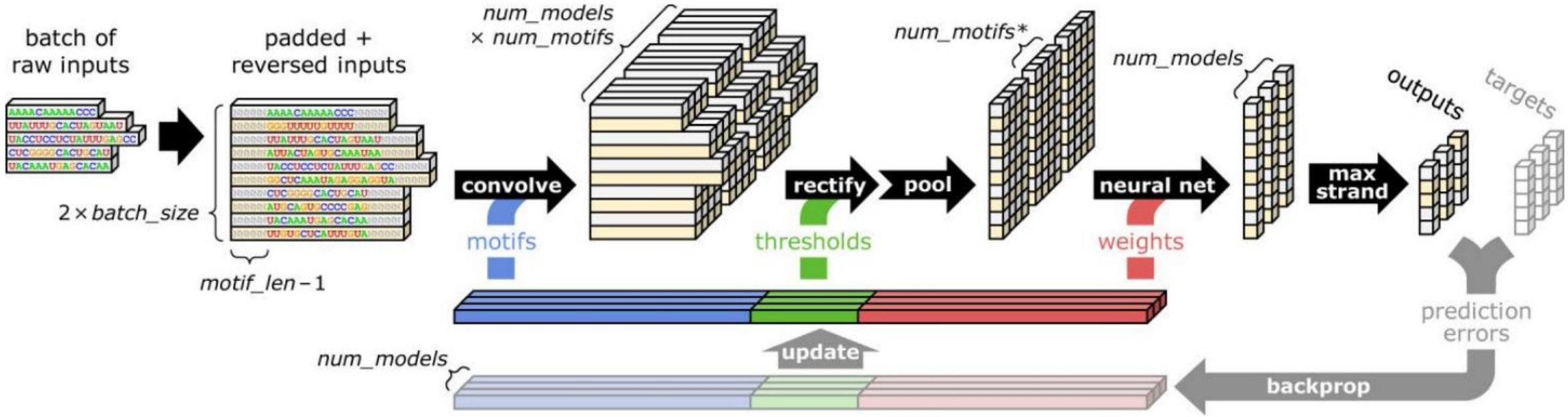

- Convolution
  - For each sequence **S**, produces a matrix **X** where element $\mathbf{X}_{i,j}$ is essentially a score of motif detector $j$ aligned to position $i$ of padded sequence **S**.
- Rectification
  - Calculate $\max(0, \mathbf{X}_{i,j} - b_j)$ for each $\mathbf{X}_{i,j}$ using threshold $b_j$.



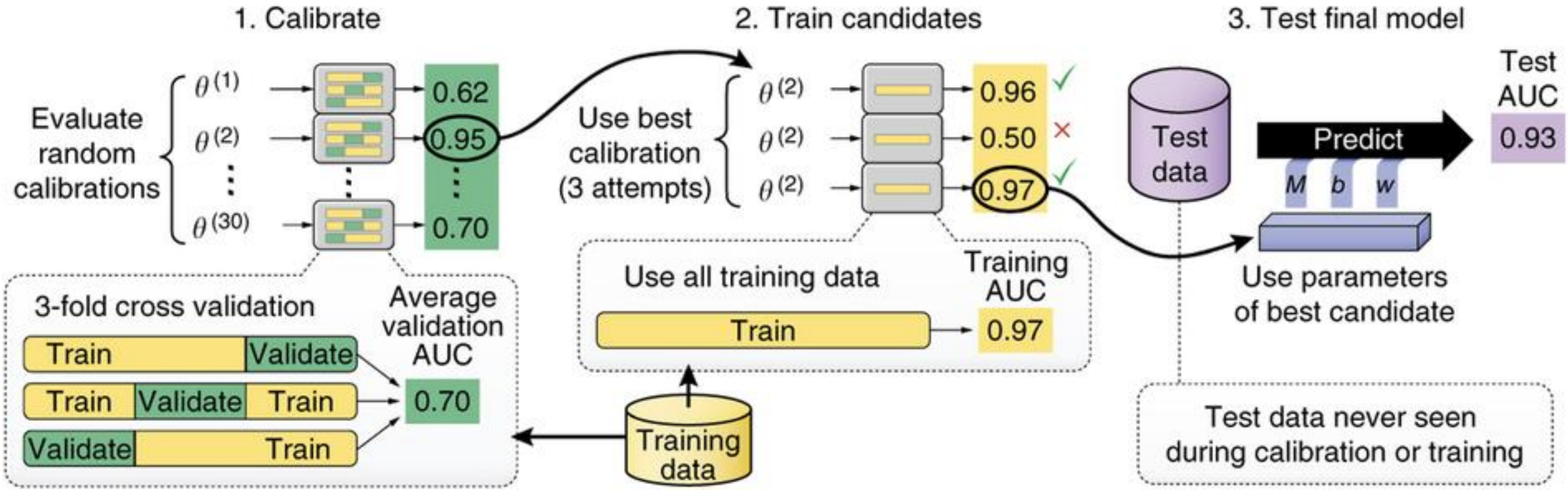

- Pooling
  - Either perform max pooling **or** max and average pooling.
  - i.e. for each column(motif detector score) ***j***, calculate max($X_{1,j}$, …, $X_{n,j}$).
  - ”RNA-binding protein models tended to benefit from knowing the average response of a motif detector within the sequence”.




- Neural network
  - Transform feature vector into a scalar output score *f* using weights *W*.
  - Can contain a hidden layer (user's choice, may improve performance).
  - Score indicates how well the input sequence matches the trained model (binding score).




- Neural network with dropout
  - Occasionally ”drop out” values in random nodes by setting them to zero.
  - Strong regularization effect.
- Parameters updated by using back-propagation with gradient descent
- Loss function is either
  - MSE (when training a model to predict microarray binding affinity measurements) or
  - Negative log-likelihood (for ChiP and SELEX data where response is binary).




- The overall training procedure used
  - 30 models with different parameters trained and evaluated using 3-fold CV.
  - Best model according to AUC trained again with same parameters using all the training data.
  - Best one chosen according to AUC.

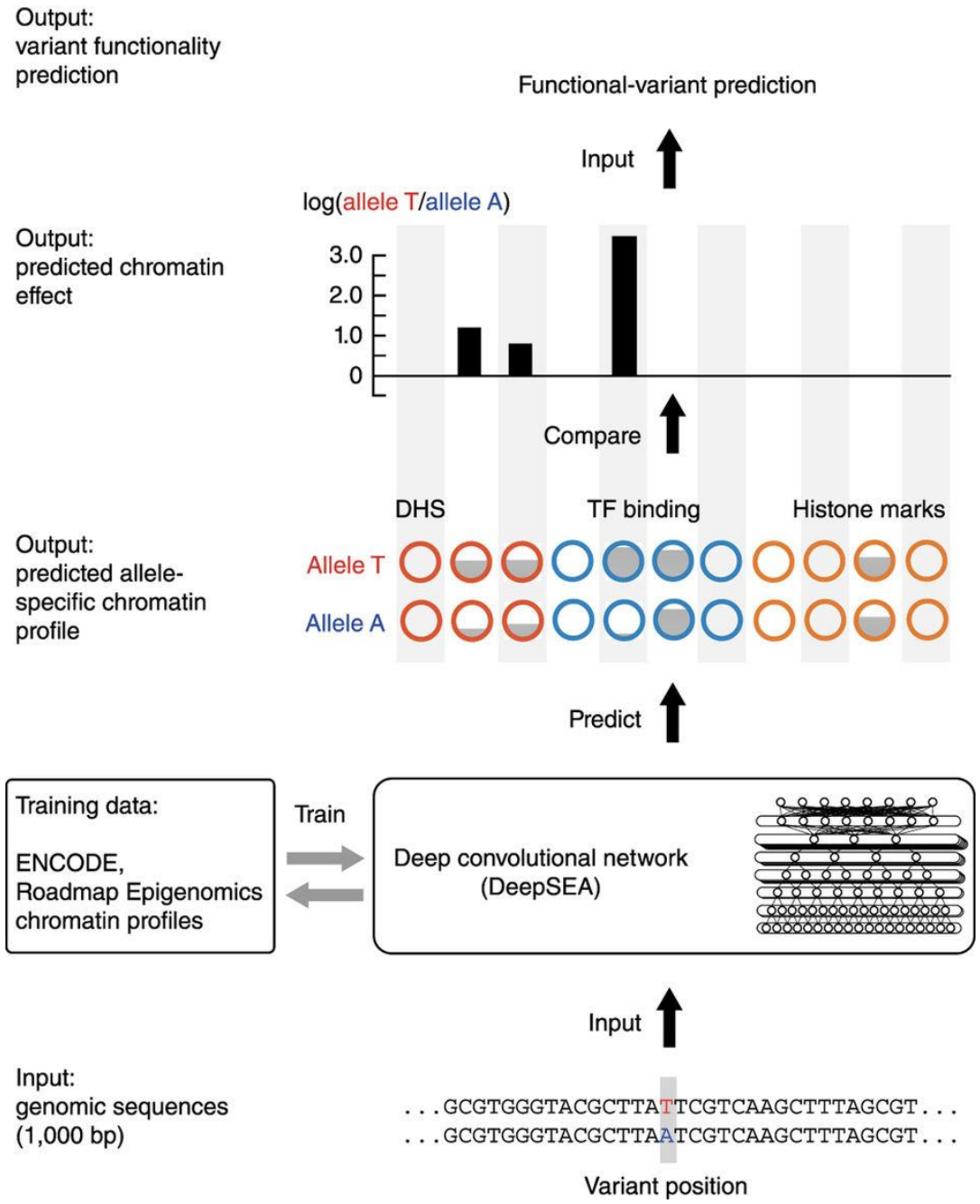

# DeepSea & DanQ

# DeepSea

- Predict noncoding-variant effects *de novo* given DNA sequence.
- Predicts chromatin effects of sequence alterations with single-nucleotide sensitivity.
- Zhou & Troyanskaya 2015, Nature Methods

# Deepsea

- Model trained with 1000bp sequences.
- 4,4 million sequences.
- 919 predictors.
- Multitask model.
- Still able to dig out sequence motifs from the network.



# DeepSea

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%

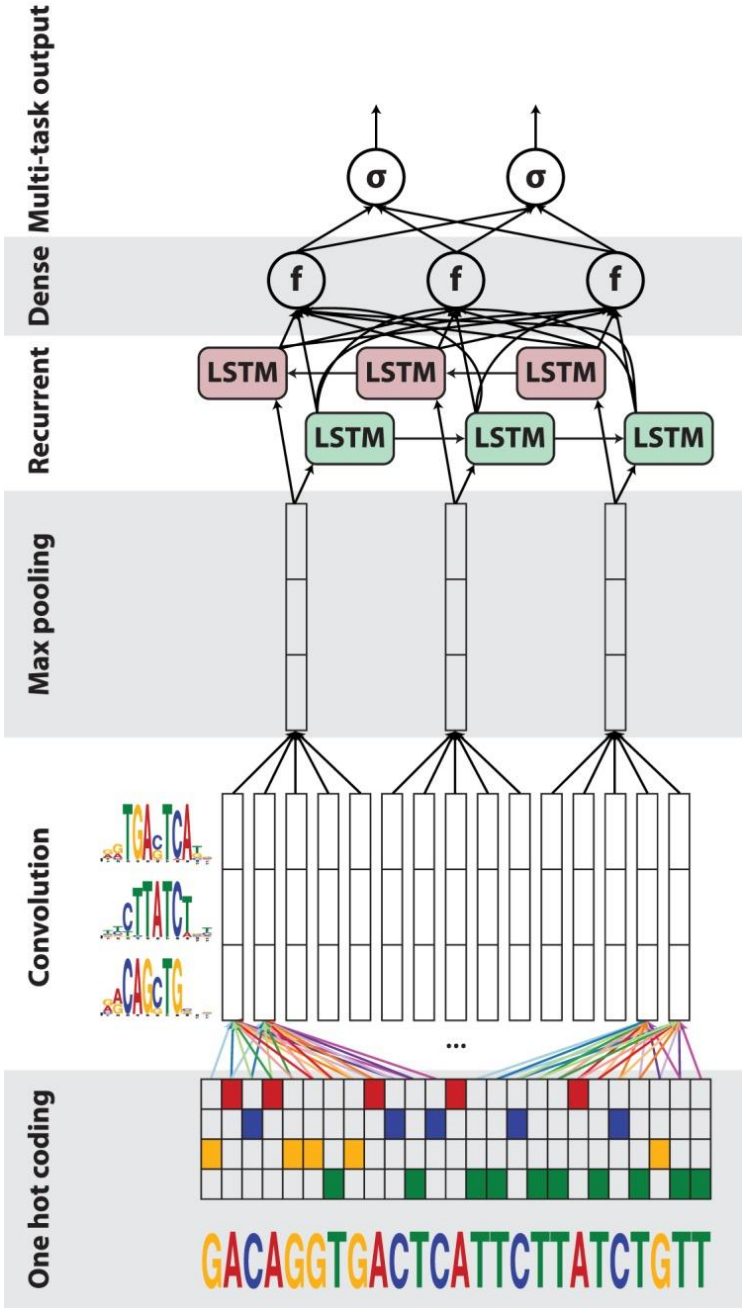
L2 regularization ($\lambda_1$): 5e-07
L1 sparsity ($\lambda_2$): 1e-08
Max kernel norm ($\lambda_3$): 0.9

# DanQ

- A hybrid framework that combines CNNs and BLSTMs.
- Same goal as DeepSea.
- The first layers of the DanQ model are designed to scan sequences for motif sites through convolution filtering.
- One convolution and max pooling layer is followed by a bi-directional long short-term memory (LSTM) layer.
- Multi-task model.
- Quang & Xie 2016, Nucleic Acids Research

# DanQ

- Convolution followed with max pooling.
- Bi-directional LSTM followed by dense layer of ReLUs (Rectified Linear Unit).
- Multi-task sigmoid output.
- Code example.



# DeepSea & DanQ

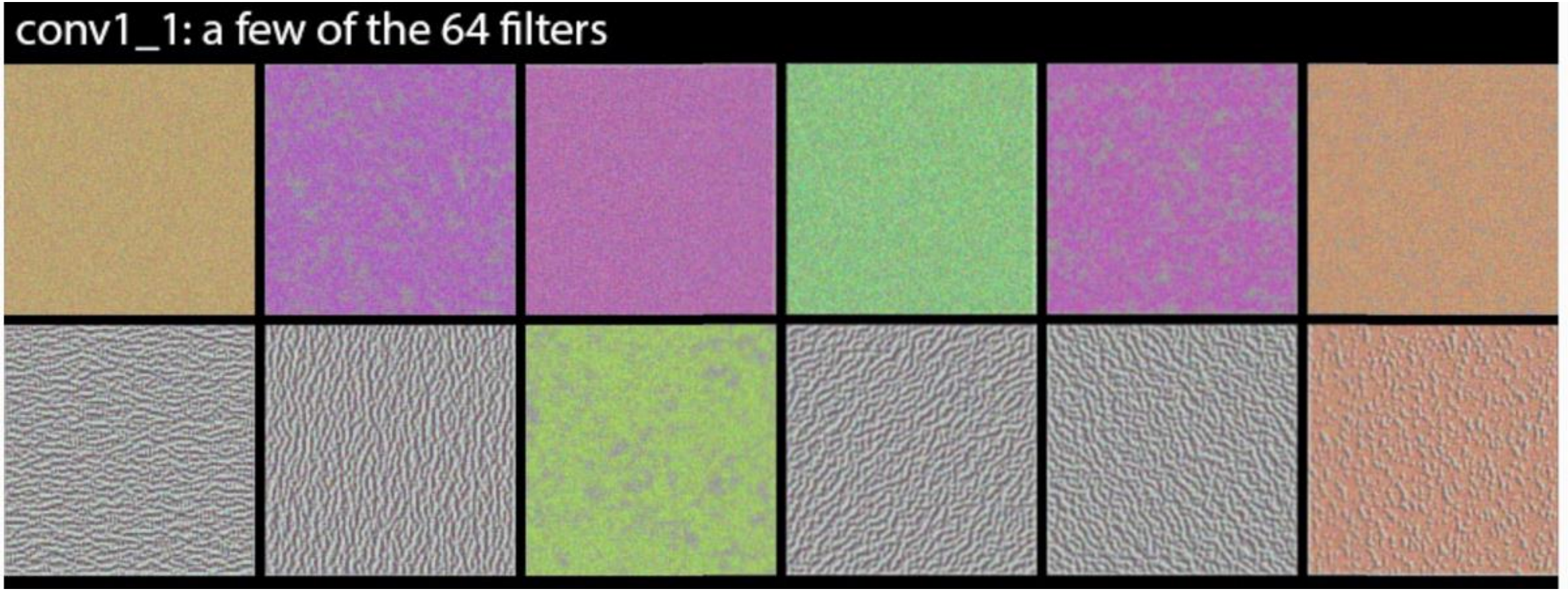
- Common components:

1. One-hot encoded inputs.
2. Convolution layer.
3. Max pooling layer.
4. Dropout regularization.
5. Linear layer

# DeepSea & DanQ

What do the filters learn?

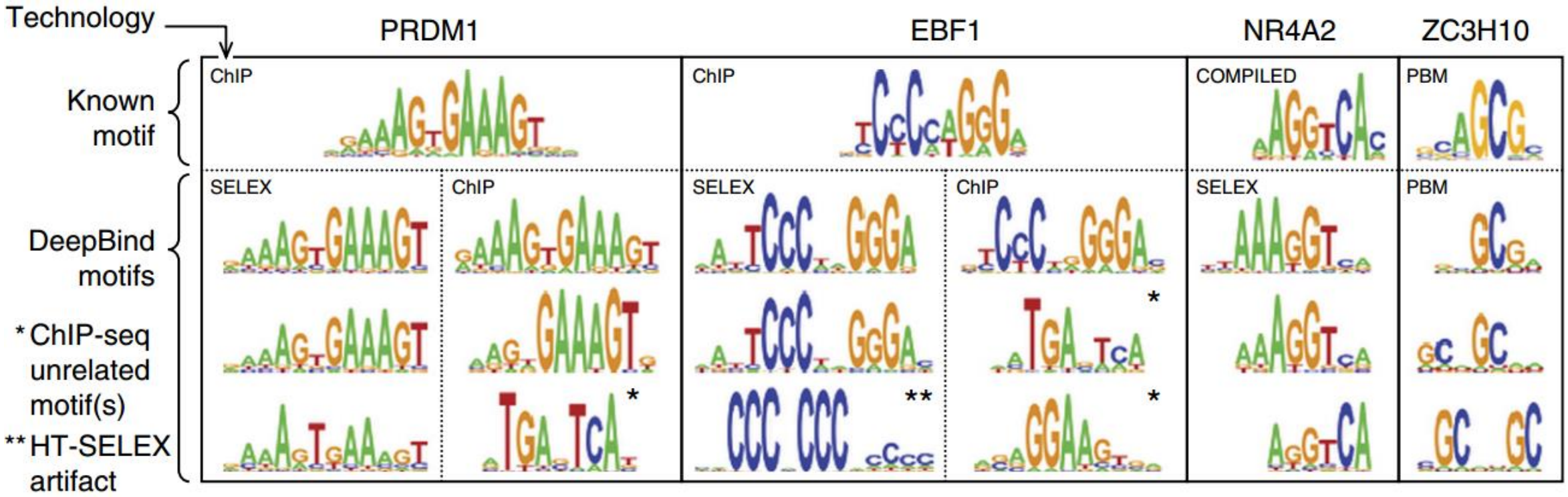
Images…



# DeepSea & DanQ

What do the filters learn?

DNA sequences…



# What else?

- Predicting changes in biological function by ”computational mutation scanning”.
- Finding an input that maximizes a specific class.