

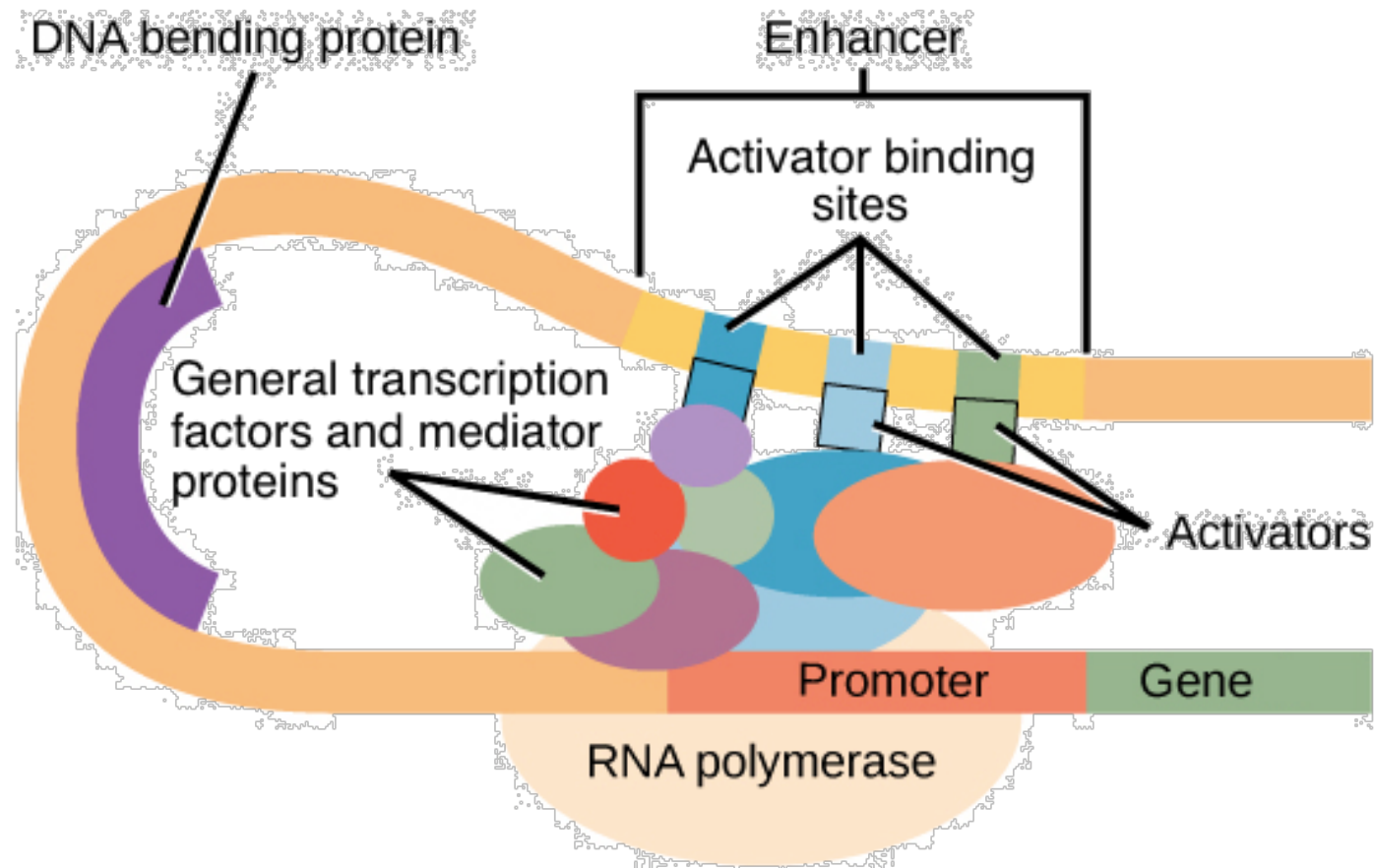
Convolutional neural network architectures for predicting DNA-protein binding

HAOYANG ZENG

GIFFORD LAB, MIT

JULY 11, 2016

DNA-protein binding is essential to cellular function



Traditional DNA-protein binding models

AAGTGT
TAATGT
AATTGT
AATTGA
ATCTGT
AATTGT
TGTTGT
AAATGA

Input

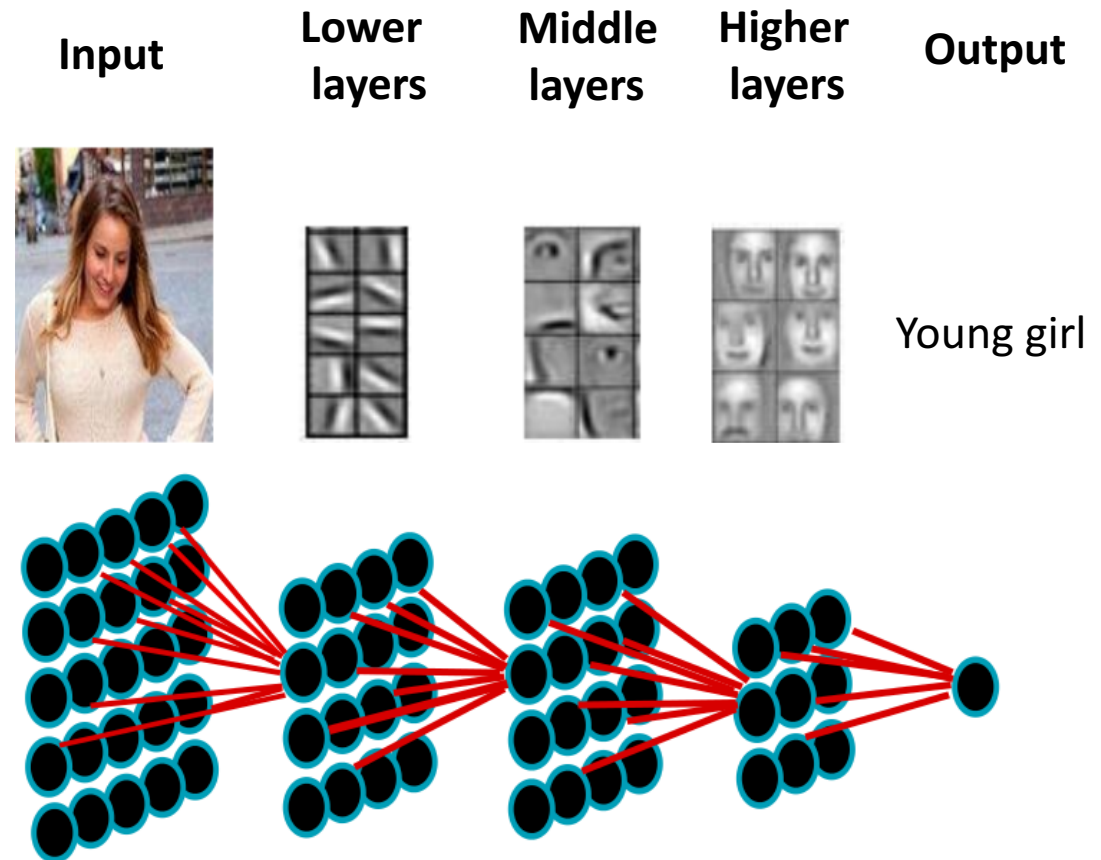
Revolutionized by deep learning !

0.25	0.01	0.25
0.13	0.01	0.01
0.13	0.96	0.01
0.49	0.01	0.73

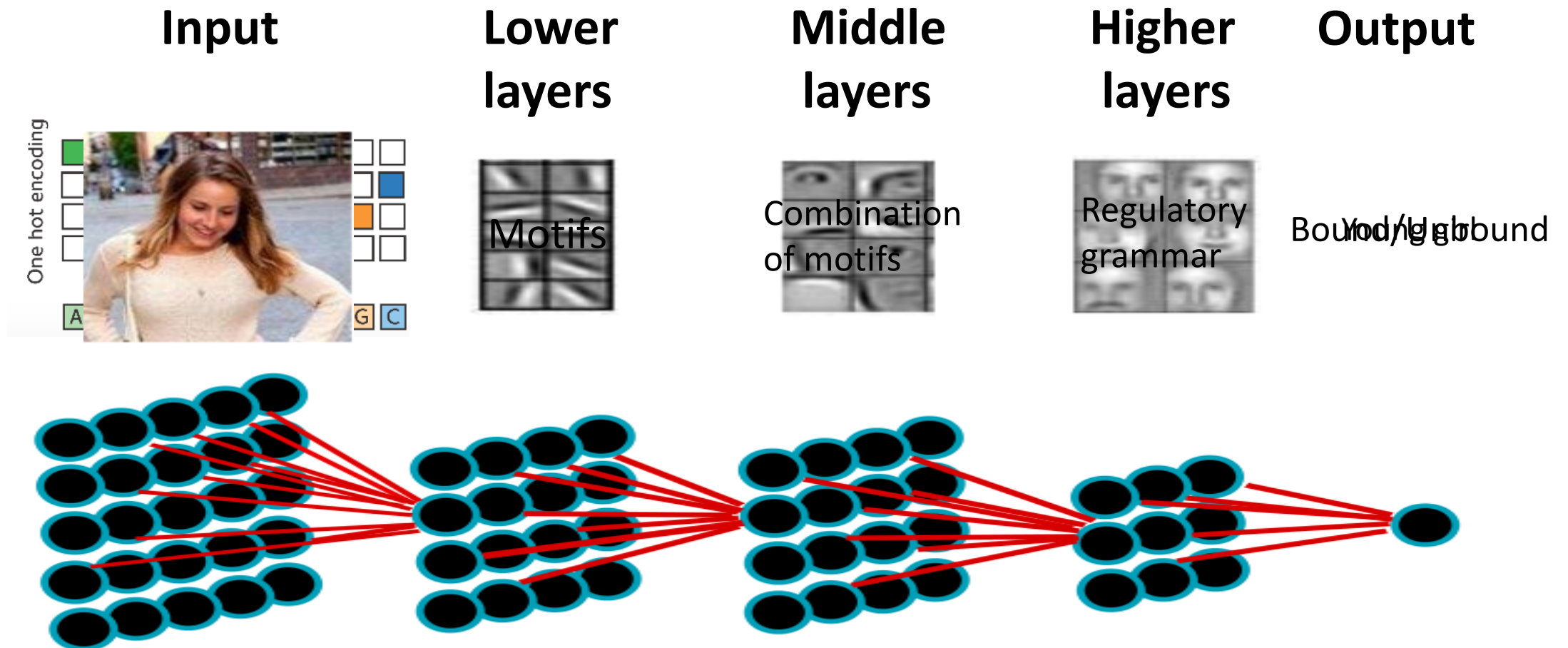
quencies

What is “Deep Learning” ?

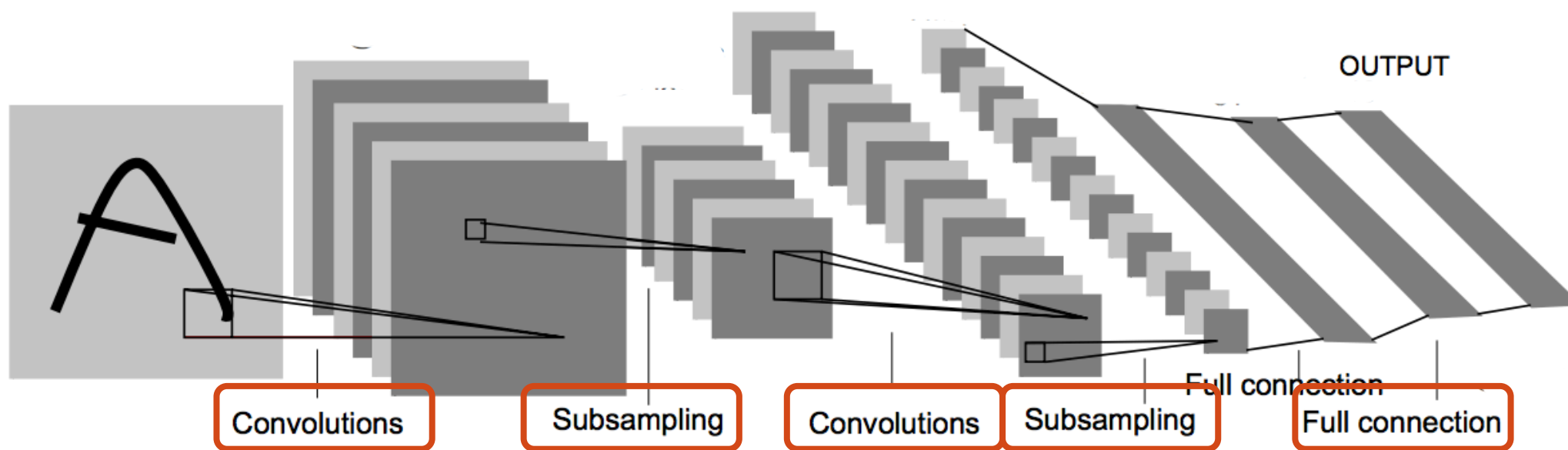
- Layered stack of neural networks
- Each layer automatically learns features of different levels
- Produces a highly non-linear mapping from input to output given sufficient training data
- Highly scalable with GPUs



How deep learning works on genomics

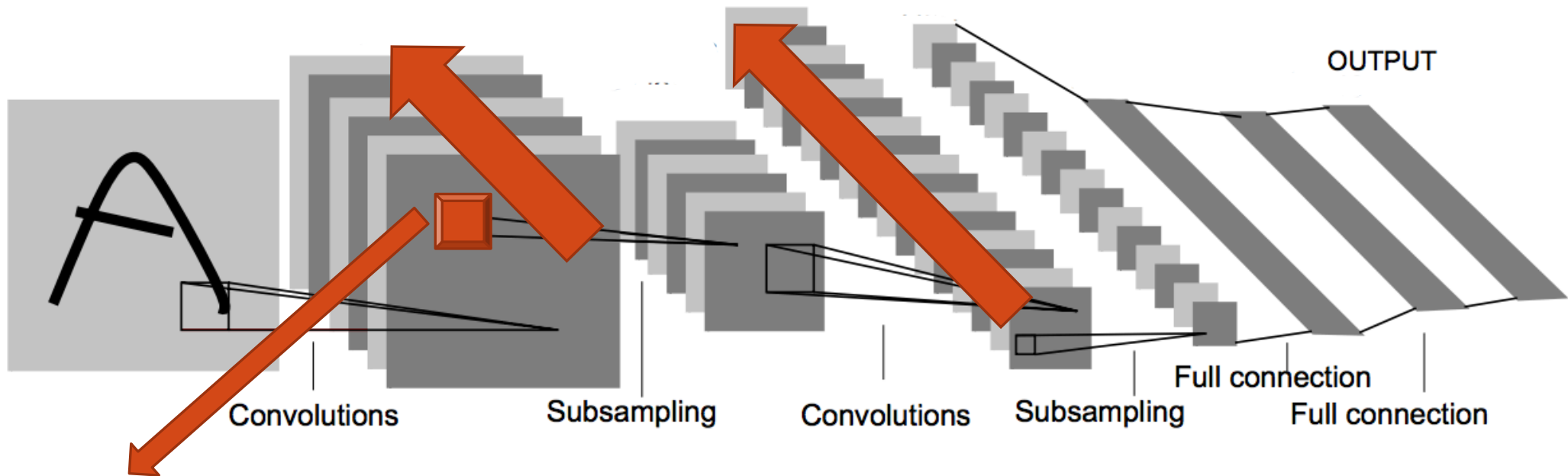


What's a convolutional neural network (CNN) ?



What's a convolutional neural network (CNN) ?

More convolution kernels better capture feature diversity



Smaller pooling window retains more location information

More layers can capture higher level features

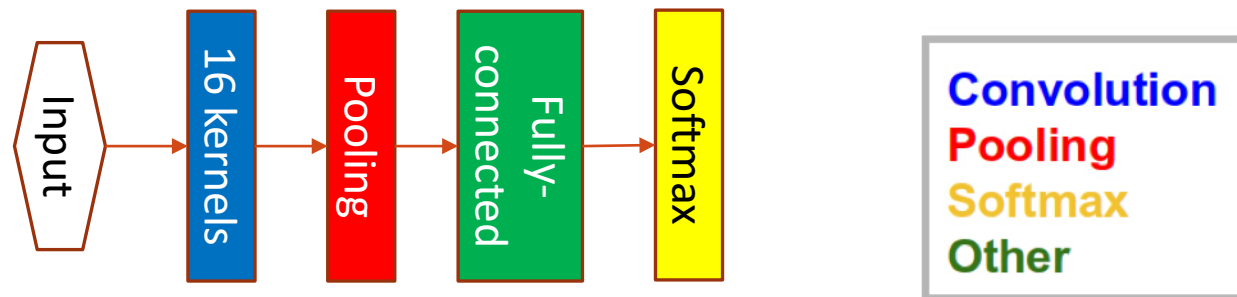
CNNs can outperform conventional approaches in modeling DNA-protein binding

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi^{1,2,6}, Andrew Delong^{1,6}, Matthew T Weirauch³⁻⁵ & Brendan J Frey¹⁻³

DeepBind (2015):

One convolutional layer with 16 kernels, maximum pooling window

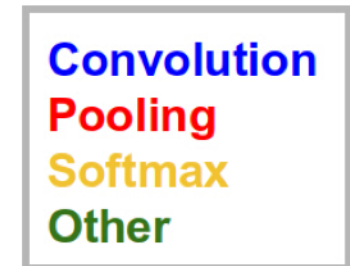
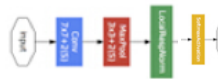


DeepBind is “shallow learning” compared with other CNNs

GoogLeNet^[1]
(Computer Vision)



DeepBind



[1] Szegedy et al. Going Deeper with Convolutions.

Open questions about deep learning for genomics

- What architectures work best to model DNA-protein binding?
- How “deep” should a network be?
- What components of the network contribute most to overall performance?
- Is the optimum network design specific to the task / experiment / TF?

Our approach

- We developed a framework to systematically benchmark CNN architectures on genomics tasks
- We analyzed the contribution of different network components
- We explored if the optimum architecture is task-specific
- We evaluated training data requirements

Systematic benchmarking is important

- Task should be meaningful
 - *Real vs. artificial sequences (DeepBind): motif discovery*
 - Simple
 - Learn motif from similar nucleotide background
 - Not generalizable to classify real bound sequences

Systematic benchmarking is important

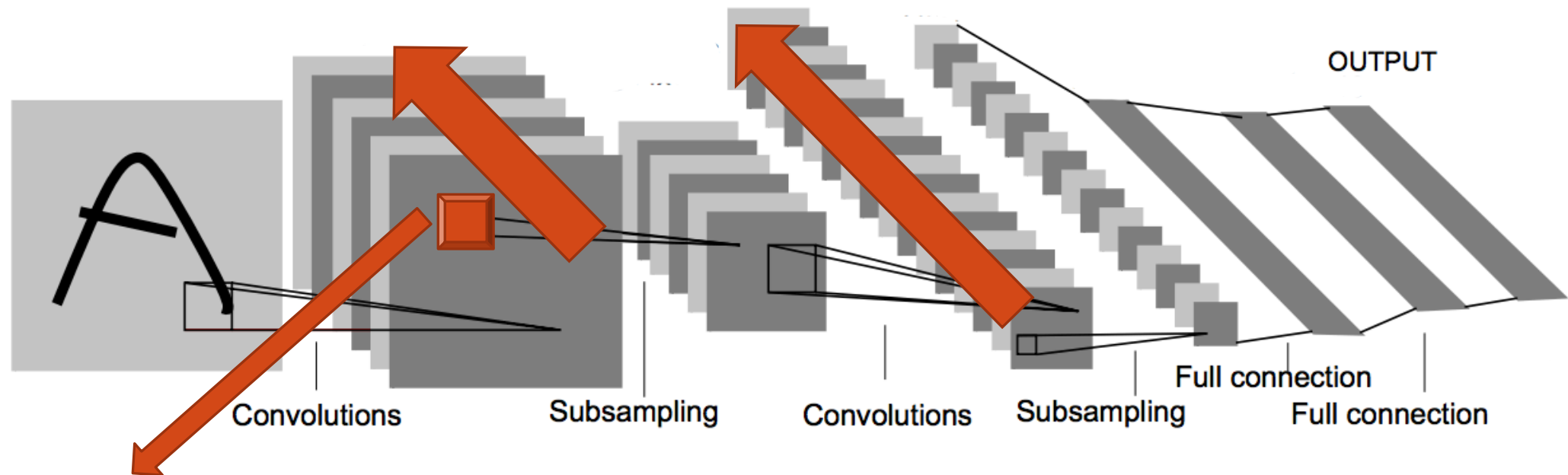
- Task should be meaningful
 - *Real vs. artificial sequences (DeepBind): motif discovery*
 - *Bound motif vs. unbound motif: motif occupancy*
 - Hard
 - Forces the model to learn better and higher-level sequence determinants

Systematic benchmarking is important

- Task should be meaningful
- Balance the number of positive and negative samples
- Control any artificial bias, location of the motif in the sample
- Conclusion should be the consensus across diverse TF ChIP-seq experiments (we used 690 from ENCODE)

CNNs have three important architectural dimensions to vary







More convolution kernels better capture feature diversity



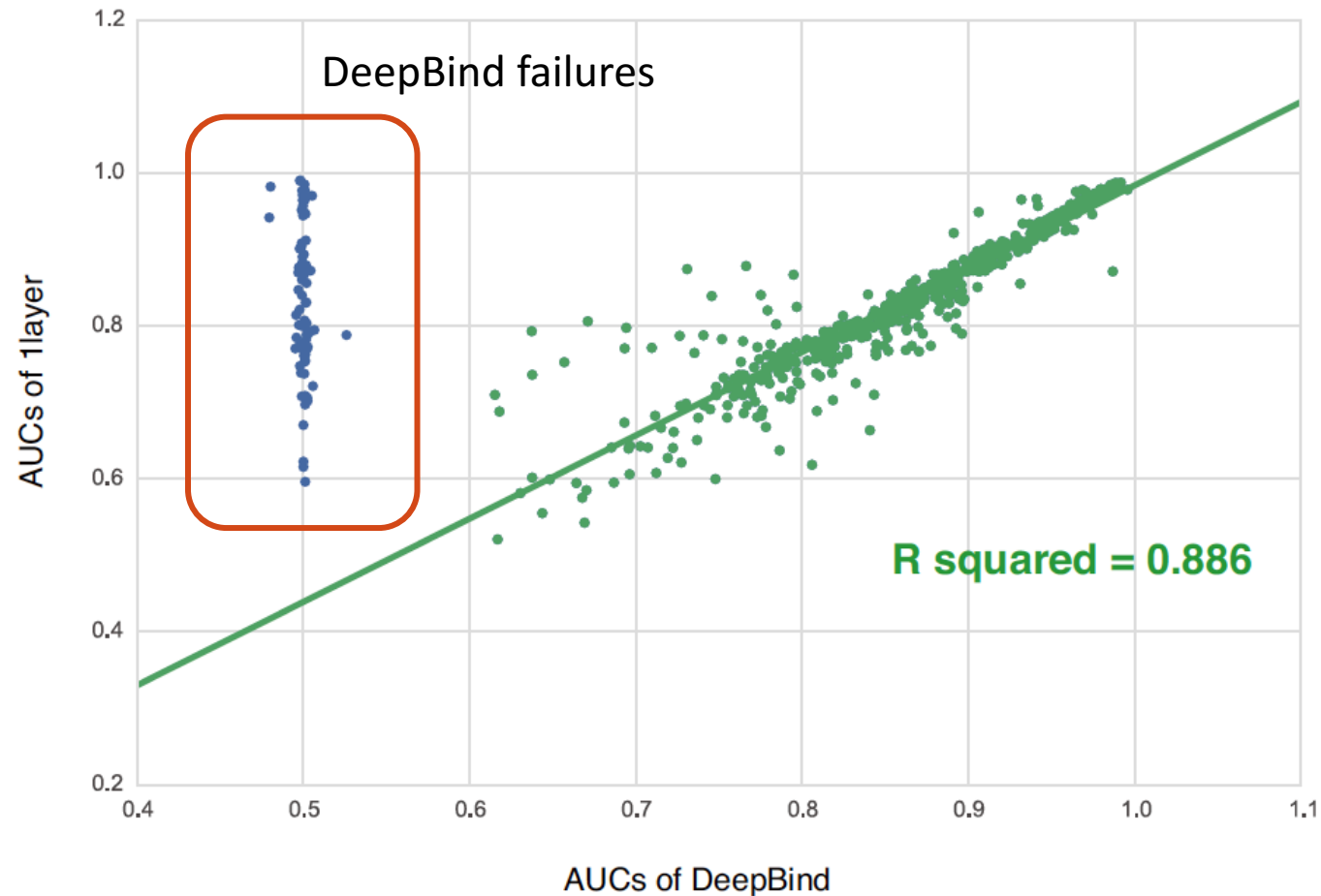
Smaller pooling window retains more location information

More layers can capture higher level features

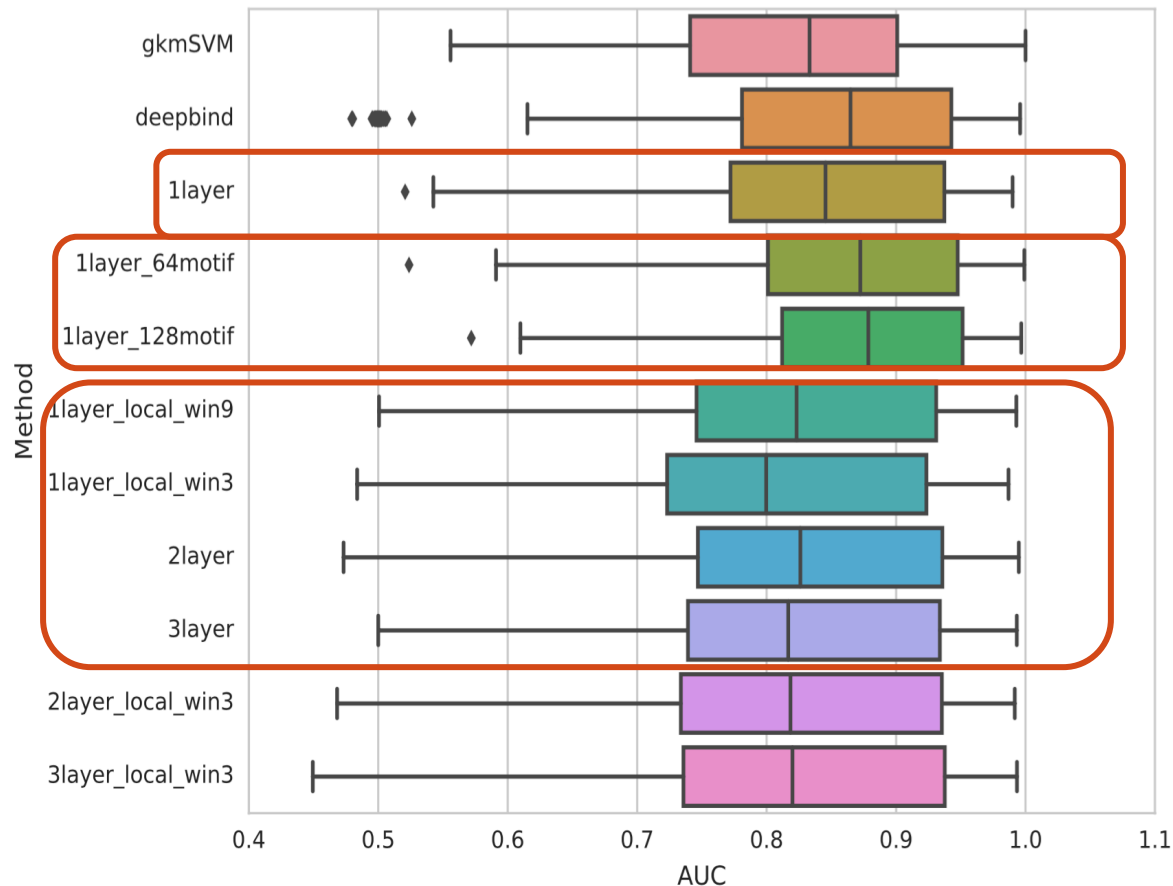
CNN architectures compared

Our Name	More Conv. Kernels	Deeper	Smaller pooling size
1layer (DeepBind)	-	-	-
1layer_64motif		-	-
1layer_128motif	 	-	-
1layer_local_win9	-	-	
1layer_local_win3	-	-	 
2layer	-		-
3layer	-	 	-
2layer_local_win3	-		 
3layer_local_win3	-	 	 

Baseline model reproduces DeepBind



Simple models are best for a **motif discovery task**



baseline

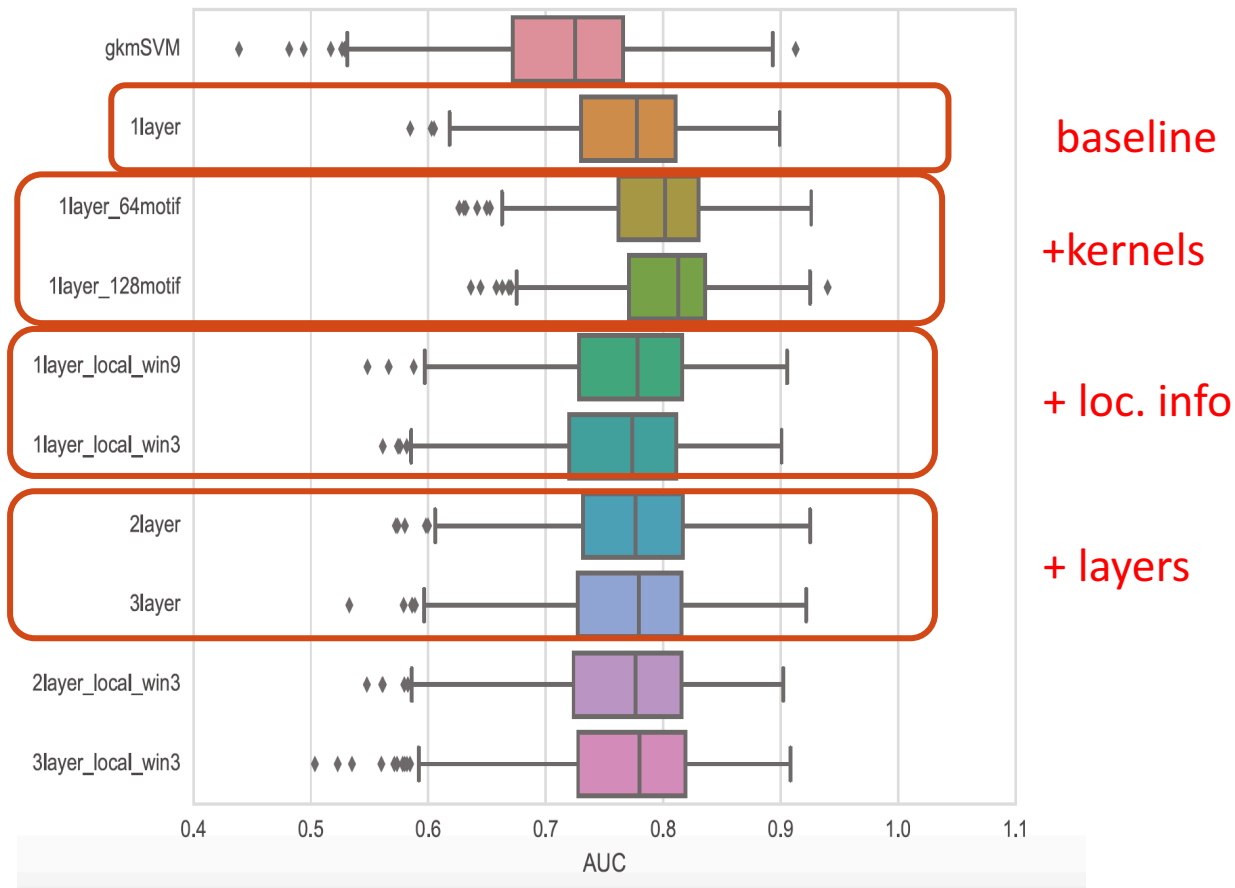
+kernels

+ loc. info

+ layers

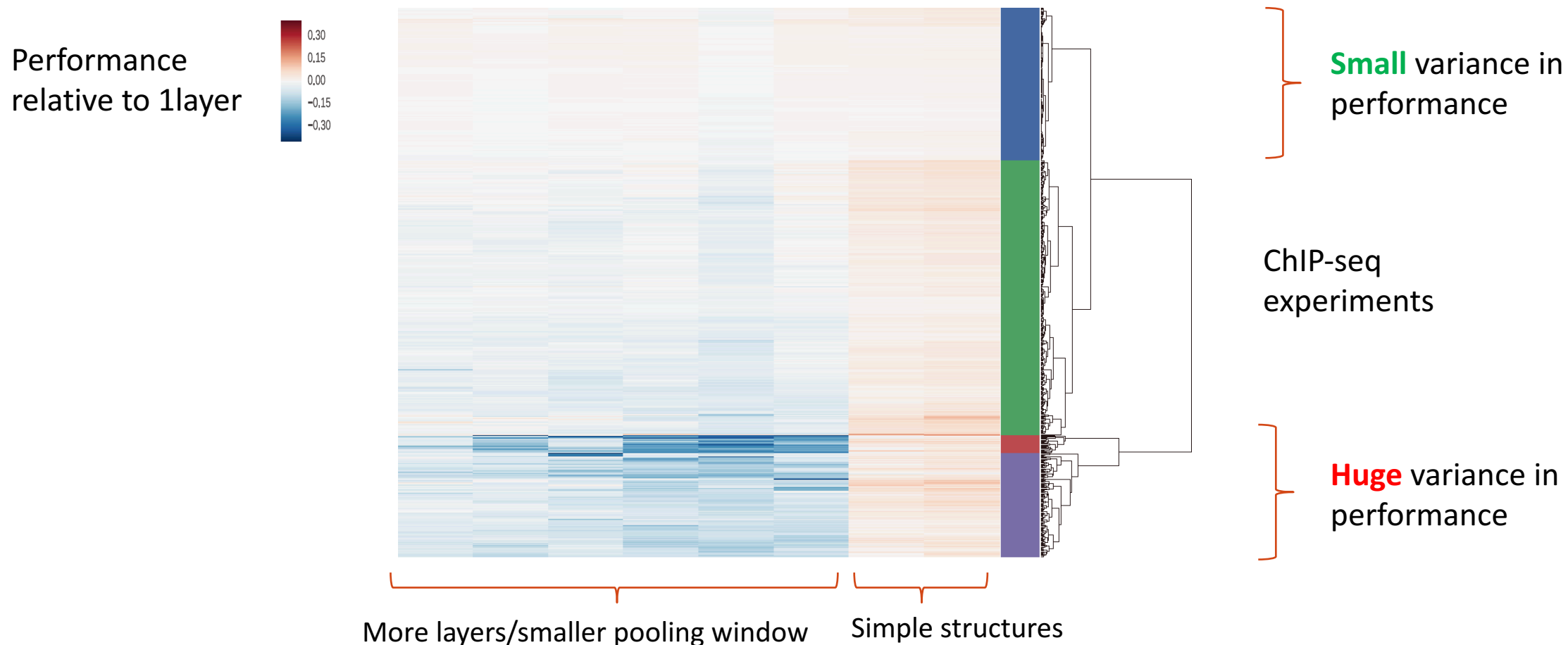
- More convolutional kernels helps model motif diversity
- Smaller pooling size, more layers monotonically decrease performance
 - possibly because most determinants are low-level (motifs) and position-independent

Depth improves performance in a motif occupancy task

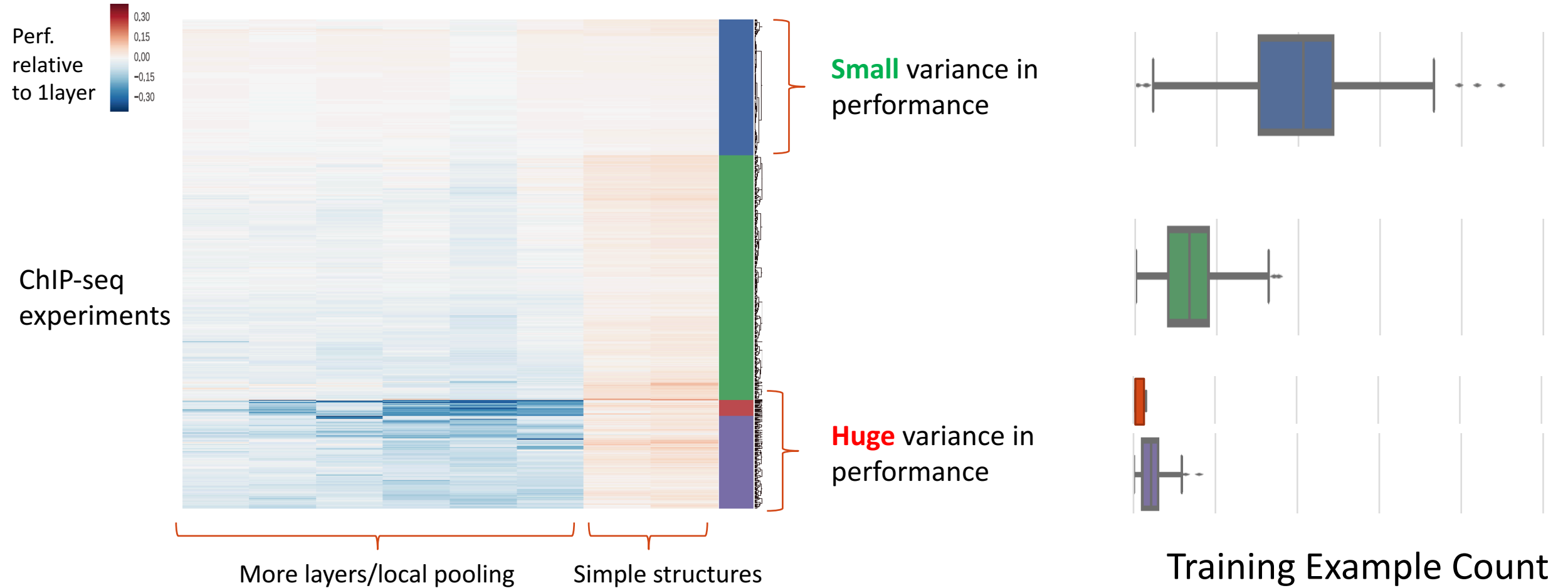


- AUC decreases for all architectures
- More convolutional kernels help model the motif diversity
- Smaller pooling size slightly decreases the performance
- Deeper networks have slightly better performance
 - There are more high-level determinants that can be better modeled by deeper layers, consistent with the task design

Observed performance is experiment-specific

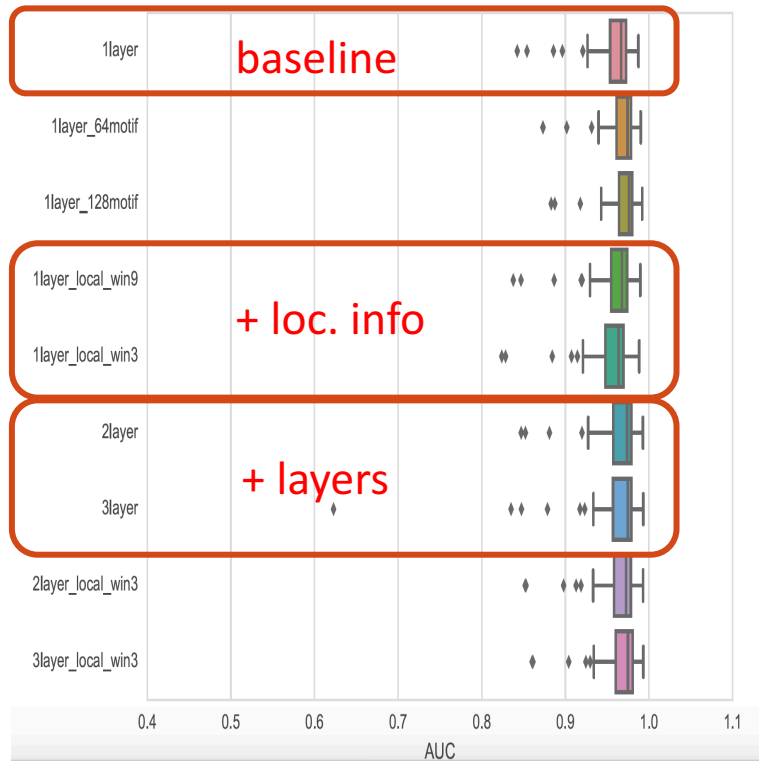


More complex networks require more training data

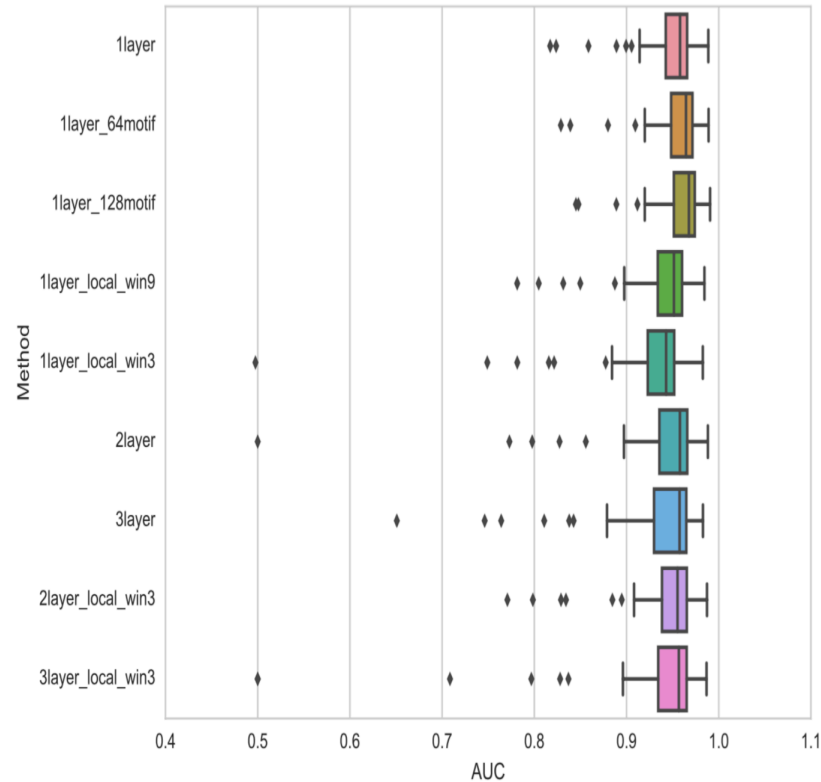


Variance increases with fewer training examples

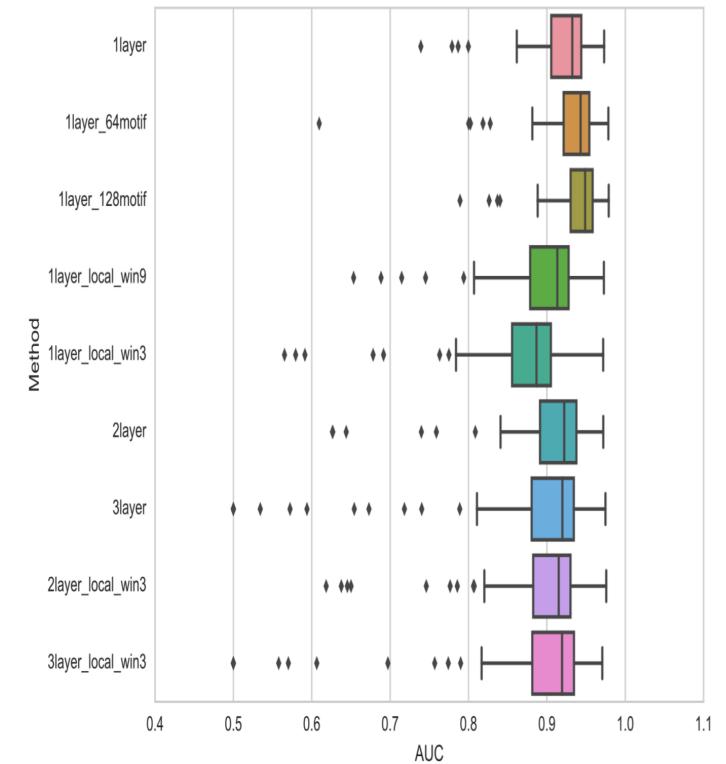
Performance on motif discovery task



80,000 training examples



20,000 training examples



5,000 training examples

TF family marginally contributes to variance

On training size controlled datasets, we performed an ANOVA test on the effect of architecture and TF family on the performance (AUC)

- Model architecture is always the only factor with significant effect
- The effect size of TF family is always 50 times less than that of the model architecture

Key results of our work

- An open source platform to deploy and compare deep learning models in the cloud (<http://cnn.csail.mit.edu>)
- A systematic analysis of CNN structures for genomics:
 - A sufficient number of convolutional kernels is essential
 - Deeper models are needed when high-level determinants exist
 - Pattern position information is not helpful for the tasks we considered
- Complex CNNs demand more training data to work well
- TF family-specific performance not observed

Observations

- CNNs outperform conventional methods with the right structure
- The optimum structure is different from that in computer vision
- Different biological tasks and data yield different conclusions
- Understanding the problem at hand and comparing different structures is important to design a good CNN model for biology applications (<http://cnn.csail.mit.edu>)

Acknowledgments



Prof. David Gifford



Matthew Edwards



Ge Liu

