Deep MR[2]

Interrogating Deep Learning Models with Mendelian Randomization

Stephen Malina in collaboration with David Knowles and Daniel Cizin

March 3, 2020

[2] Title is a WIP.

Outline

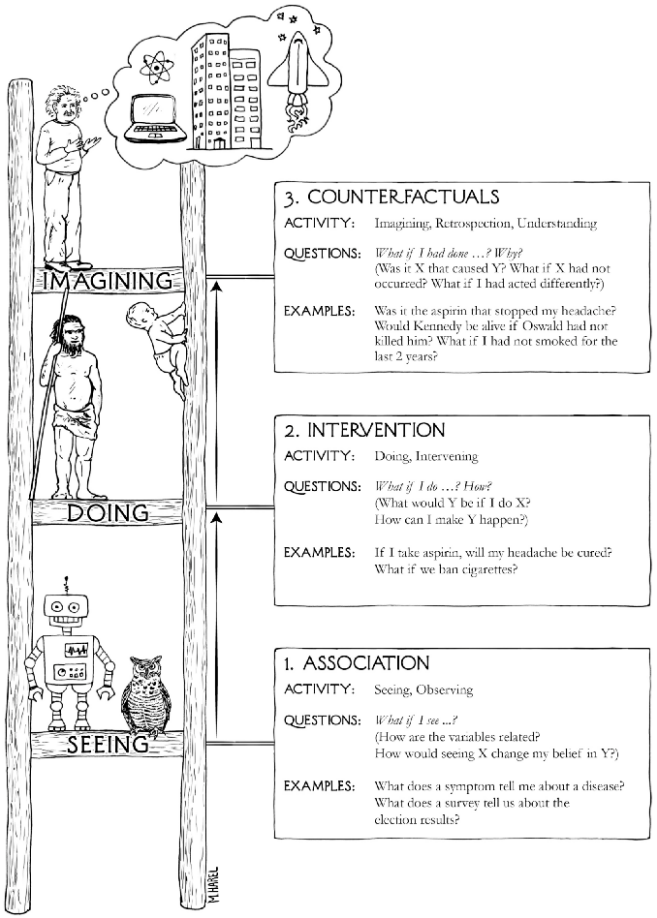
Disclaimer

I made this presentation more provocative than usual. I apologize in advance...

Motivation

ML perspective: Do deep learning models learn causal relationships?

If I train my deep learning model to jointly predict two processes one of which causes the other, will it learn the causal relationship between the two?



Biology Perspective: Microscope Models[7]

Pretend you're a biologist...

If you're a biologist, pretend you're a 'pretend biologist'.

Imagine a model of a complex process like splicing plus RBP binding that you can use to:

1. Ask questions about / verify underlying mechanisms
2. Generate hypotheses for mysterious things to look into

Without our help!

Claim

We are far from this... Why? Lots of reasons!

[7] Stolen from Chris Olah

Slicing off a piece of the problem

One way to try and better understand what our models have learned: look at whether they model the 'correct' relationships between different processes.

Recap

Clarification

All the work I've mentioned so far is great! Haven't talked about generative models, but they try to solve some of these problems.

Summary

- Complex models that work are great
- Maximizing the value we get from complex models requires better techniques for understanding what they've learned

Looking ahead

Leveraging tools from causal inference to better understand what our models have learned.

Background

Causal Inference & Mendelian Randomization

To the whiteboard!

Deep learning for DNA/RNA sequence-specific prediction

- Tons of work using CNNs/RNNs/etc. to predict biological processes from sequence
- Examples include:
  - Transcription factor (TF) binding [1, 12]
  - Chromatin accessibility (CA) [12, 4]
  - RNA binding protein (RBP) binding [11, 6]
- We'll focus on first two but method in principle applies any time we have two causally related processes we're jointly modeling

Interpretability, briefly

Two high-level approaches to understanding what sequence predictor networks have learned:

1. In-silico mutagenesis (see [12, 4])
   - Take a sequence from your test set
   - Mutate it
   - See how the predictions change
2. Gradient-based approaches (see [8, 9])
   - Make a prediction for a sequence
   - Look at gradient of the loss w.r.t. (some function of) input

Transcription factors regulate accessibility & protein production

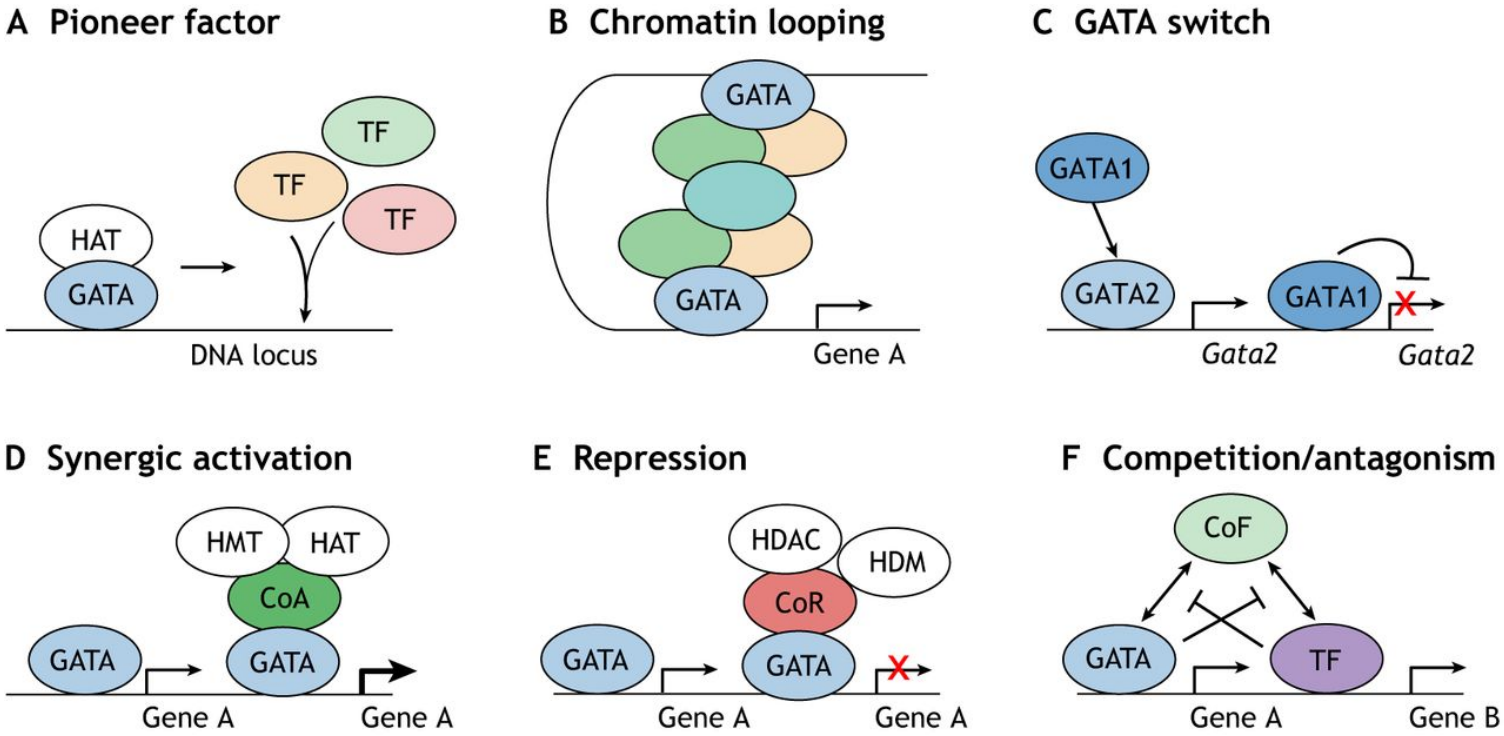


Figure: Six mechanisms through which transcription factor binding can regulate transcription[10].

Measuring transcription factor binding with ChIP-seq

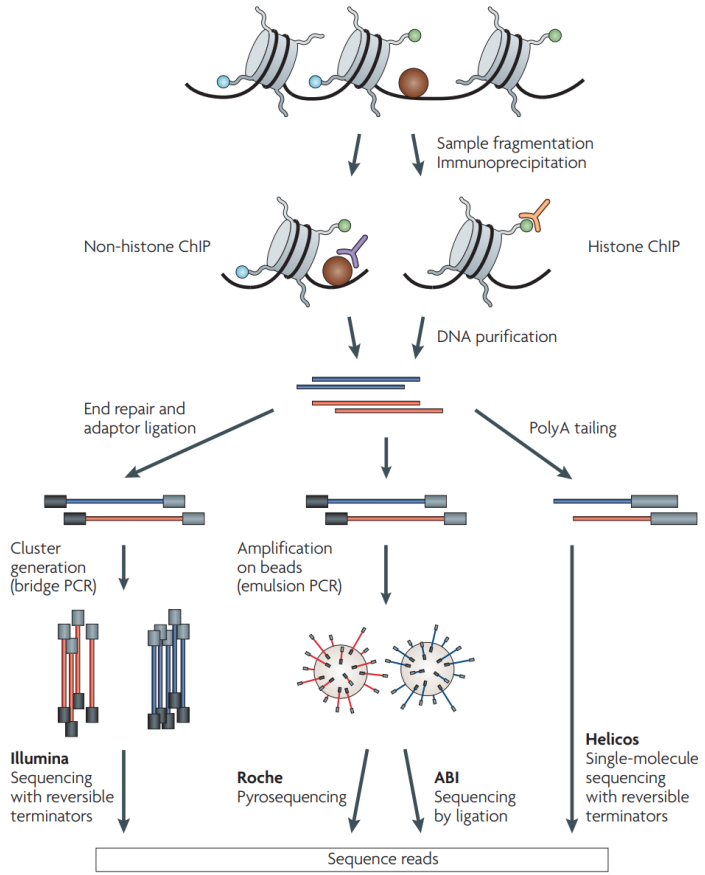


Figure: How ChIP-seq (Chromatin Immunoprecipitation sequencing) works [7].

Chromatin shape determines DNA transcription

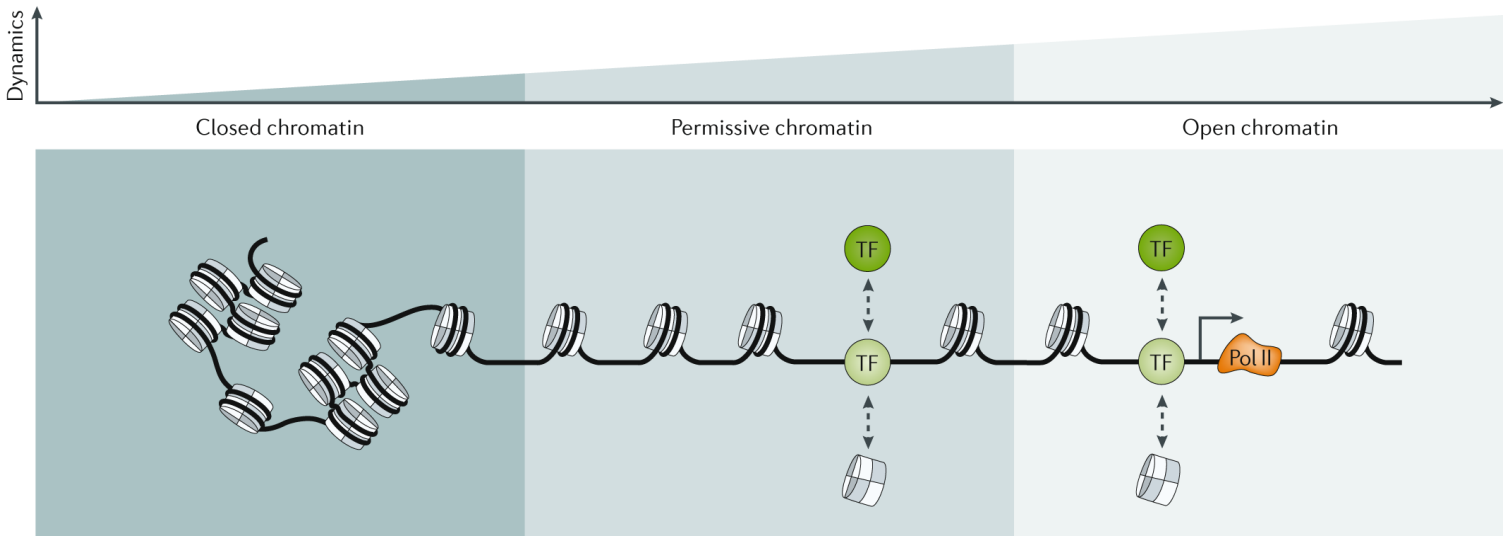


Figure: Mechanism by which chromatin accessibility regulates transcription [5].

Measuring chromatin accessibility

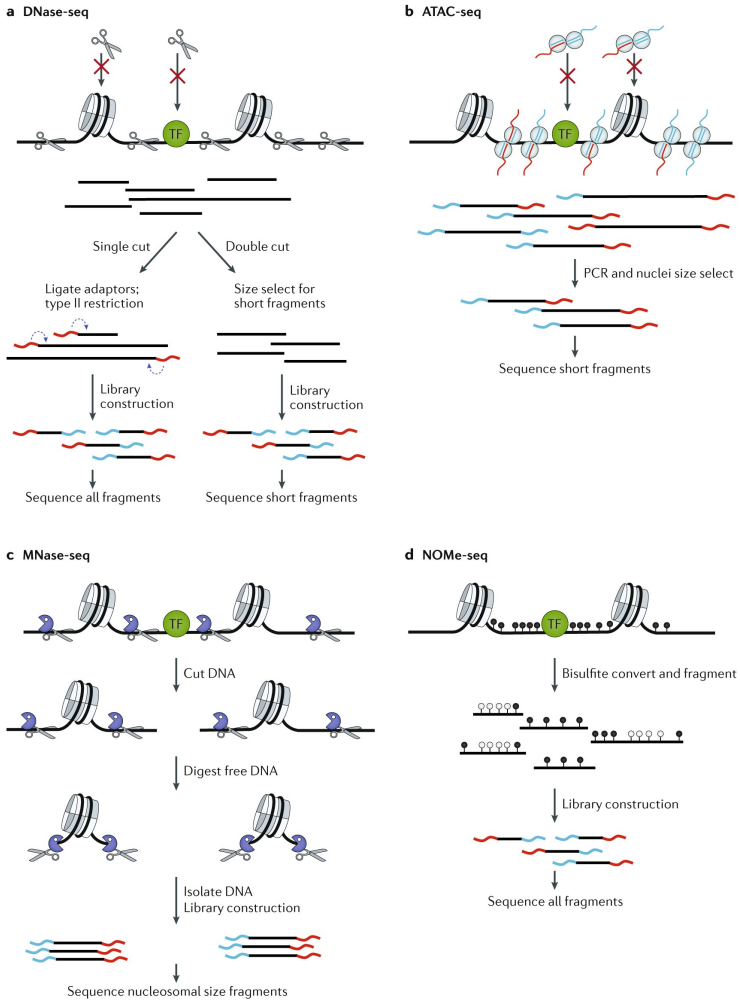


Figure: Four methods for measuring chromatin accessibility [5].

Methods

Questions we hope our method will answer

- Do jointly trained models learn[20] causal relationships?
- According to this model, does binding of a pioneer transcription factor causally influence chromatin accessibility (and vice versa)?

[20] Where 'learn' is defined in terms of whether they can generate data that reflects them.

Intuition / Assumptions

- Want to estimate whether one process (binding) causally influences another (accessibility) at the sequence region level and at the sequence population level
- If our model 'understands' causal relationships, we should be able to get it to generate data that reflects them
- "Average causal effects" is a meaningful concept at the sequence region and sequence population levels

Step-by-step Overview

1. Jointly train DL model to predict candidate cause(s) ('exposure') & effect(s) ('outcome')
2. Randomly sample sequences from held-out set and do in-silico saturation mutagenesis on each
   - Generate exposure and outcome predictions using DL model
   - Use MC dropout to generate standard errors for predictions
3. Treat difference between probabilities between each mutation and the reference sequence as effect size, analogous to a GWAS summary statistic
4. Use MR to estimate average causal effects for each sequence
5. Use meta-analysis method to aggregate causal effects across sequences

Details

To the whiteboard!

Preliminary Results

"Raw" Data

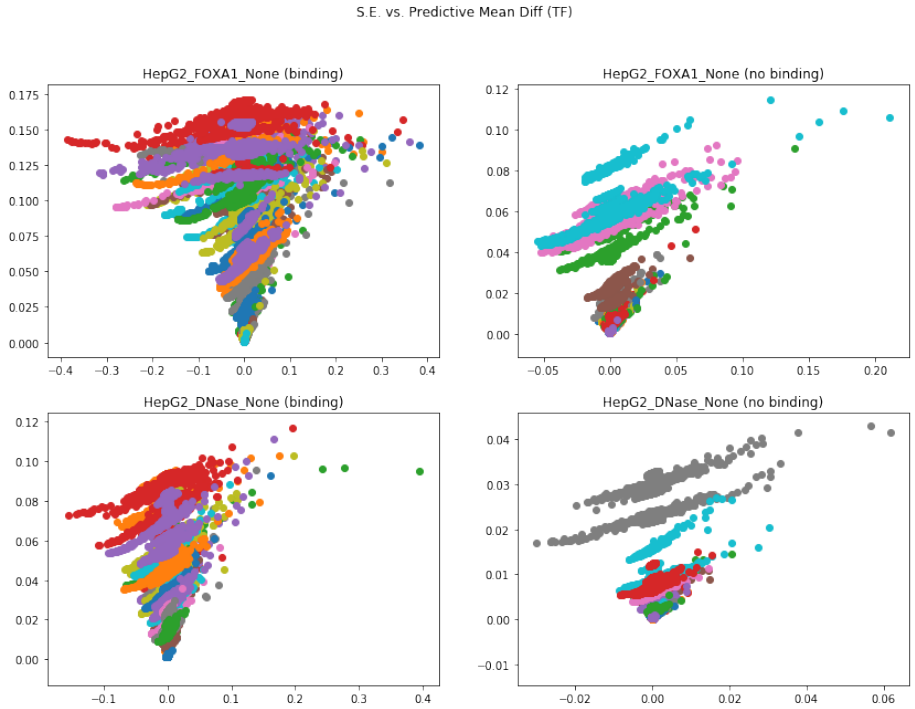


Figure: Scatter plots of standard errors as a function of predictive mean for 50 randomly sampled sequences. Two top plots represent prediction for binding and non-binding sequences respectively. Two bottom plots are split same way for chromatin accessibility.

Causal Effect Estimate Example

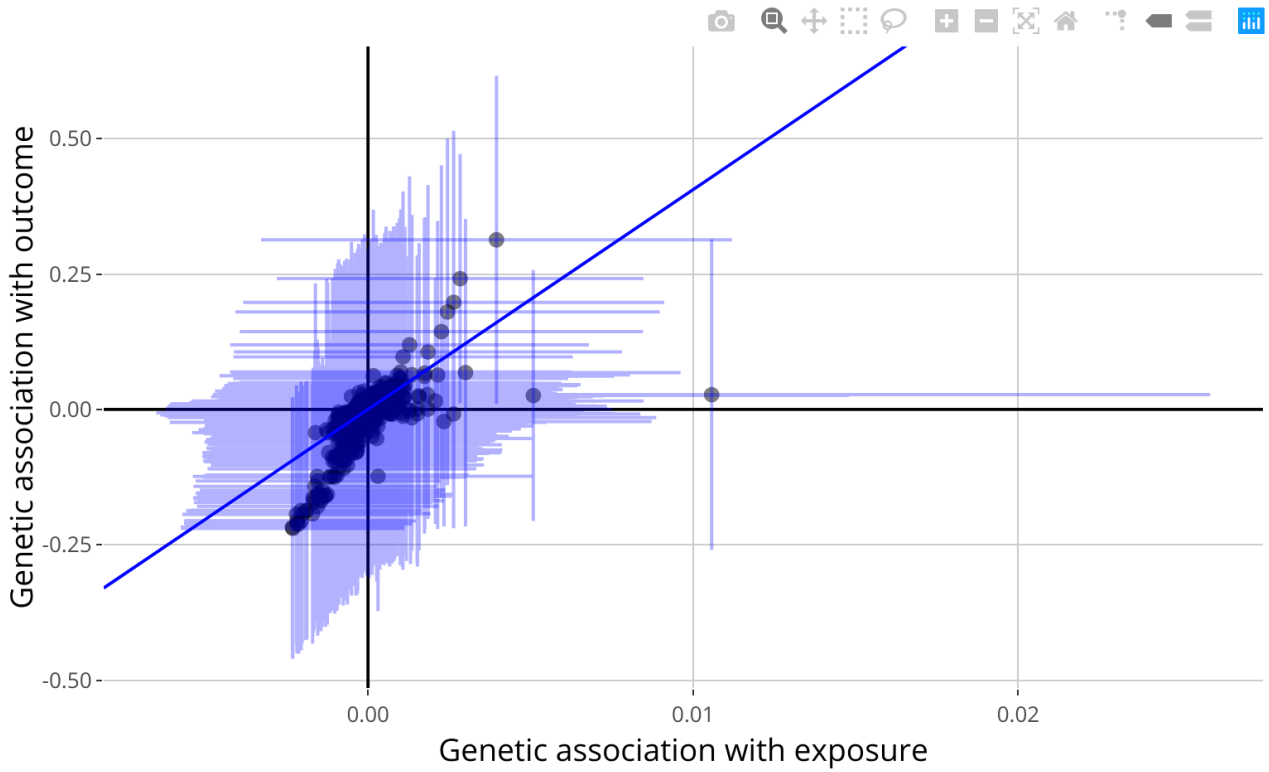


Figure: Preliminary results of running MR (Egger) for the predictions for a single sequence. The dots represent predictions for different mutation/reference pairs and the crosses standard errors.

Limitations

Our deep net is deeply uncertain

Recall the preliminary results graphs.

- Our uncertainty (standard error) is too high, overwhelming any signal
- Seems like the network may be over-estimating its uncertainty
  - **Intuition 1:** High accuracy should correlate with relatively low uncertainty
  - **Intuition 2:** Similar to a probability distribution, uncertainty should have some normalizing factor
- Can we test / fix the calibration of our uncertainty?

Our deep net is deeply mis-calibrated

- Calibration [2]:
  - Formally:

$$\Pr[\hat{y} = y \mid \hat{p} = p] = p$$

  - Semi-formally: our model assigns probability $\hat{p}$ to $(100 \times p)\%$ of samples with label 1
  - Informally: if we look at all the sequences to which our model assigns high probability of label 1 (e.g. binding), most of them actually have label 1
- Accuracy is not calibration (even with NLL loss)!
- DeepSEA is quite accurate but not very calibrated

Our deep net is deeply mis-calibrated (cont.)

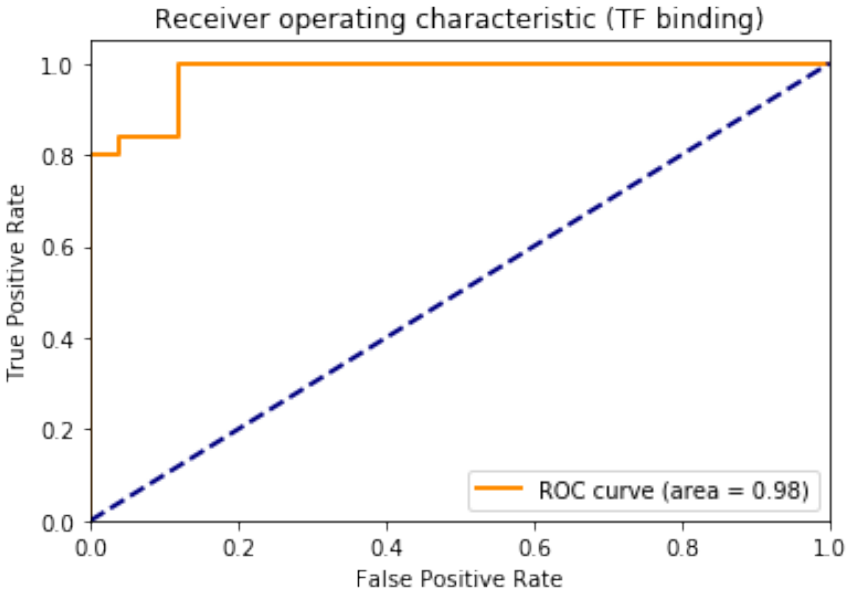


Figure: DeepSEA AUC for randomly sampled sequences. Looking good!

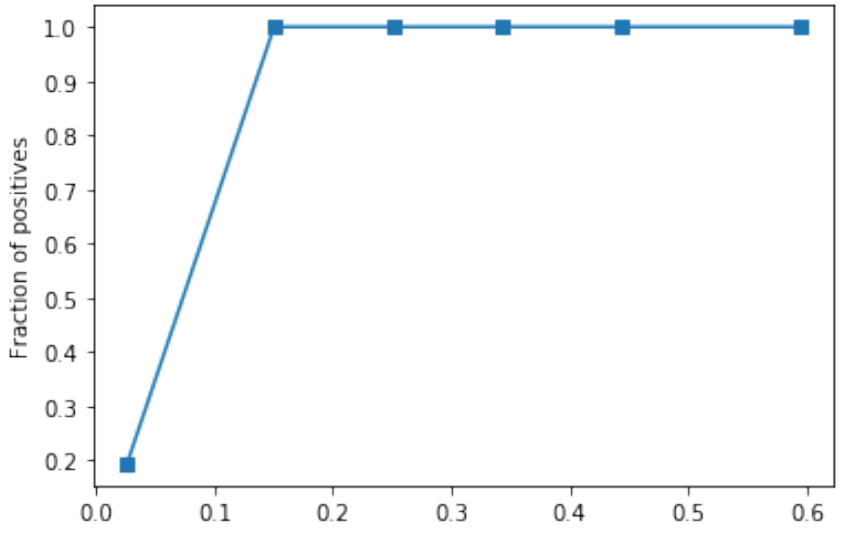


Figure: DeepSEA calibration for randomly sampled sequences. Not looking so good...

Our deep net is deeply mis-calibrated (cont.)

What's really going on here?

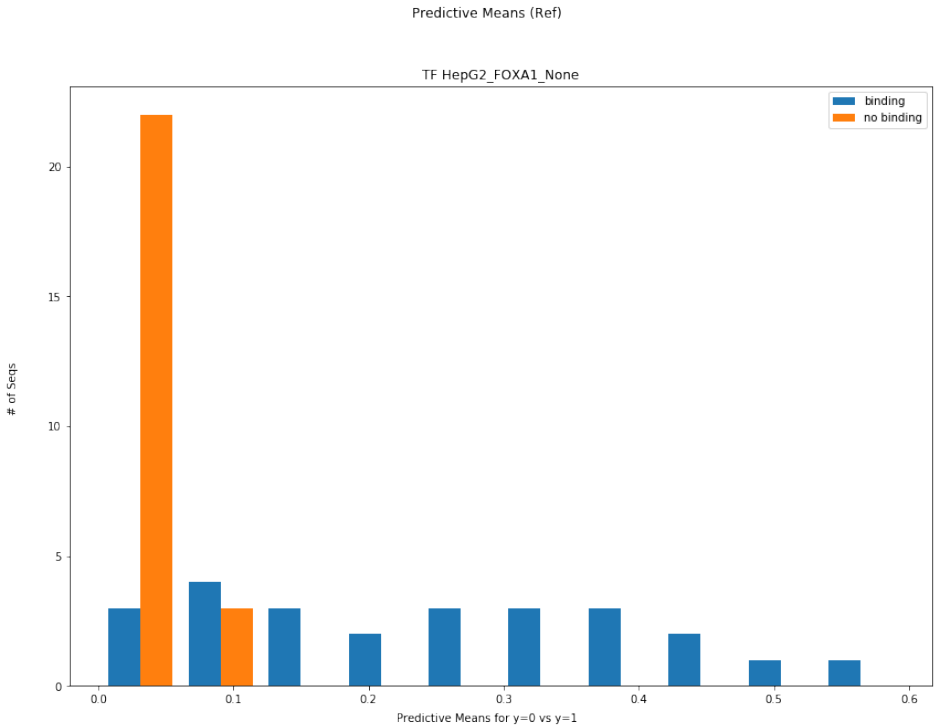


Figure: Histogram of binding probabilities. Note that overlap is low but really far to the left.

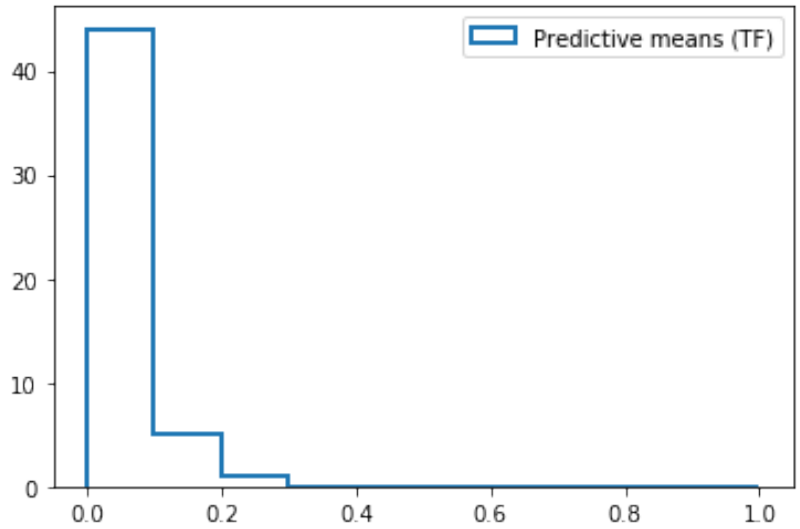


Figure: DeepSEA calibration histogram for randomly sampled sequences. Not looking so good...

Upcoming & Future Work

Near Term Work

- Measure calibration of uncertainty not just predictions
- Figure out whether we can calibrate DeepSEA without re-training it
  - Isotonic regression
  - Temperature scaling
- Bayesian MR for mode-based estimator

Long Term Work

- Improve efficiency so can run for 1000s of sequences
- Learn a causal model directly (Deep IV [3])

Conclusion

If you only remember one slide...

Goal

To go from models to mechanisms and hypotheses.

How to get there?

Methods for interrogating complex models and the **relationships** they learn at a higher level.

Our contribution

Think of understanding what a network learns through the lens of causal inference. Focus on average effects rather than individual predictions.

Thank you!

Acknowledgements

Thank you to:

- David
- Brielin
- Udai

for all of your help!

Bibliography I

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J.

Predicting the sequence specificities of dna-and rna-binding proteins by deep learning.

Nature biotechnology 33, 8 (2015), 831.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.

On calibration of modern neural networks.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M.

Deep IV: A flexible approach for counterfactual prediction.

In *Proceedings of the 34th International Conference on Machine Learning* (International Convention Centre, Sydney, Australia, 2017), D. Precup and Y. W. Teh, Eds., vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 1414–1423.

Bibliography II


Kelley, D. R., Snoek, J., and Rinn, J. L.
Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks.
Genome research 26, 7 (2016), 990–999.


Klemm, S. L., Shipony, Z., and Greenleaf, W. J.
Chromatin accessibility and the regulatory epigenome.
Nat. Rev. Genet. 20, 4 (Apr. 2019), 207–220.


Koo, P. K., Anand, P., Paul, S. B., and Eddy, S. R.
Inferring Sequence-Structure preferences of RNA-Binding proteins with convolutional residual networks.
Sept. 2018.


Park, P. J.
ChIP-seq: advantages and challenges of a maturing technology.
Nat. Rev. Genet. 10, 10 (Oct. 2009), 669–680.

Bibliography III


Shrikumar, A., Greenside, P., and Kundaje, A.
Learning important features through propagating activation differences.


Simonyan, K., Vedaldi, A., and Zisserman, A.
Deep inside convolutional networks: Visualising image classification models and saliency maps.


Tremblay, M., Sanchez-Ferras, O., and Bouchard, M.
GATA transcription factors in development and disease.
Development 145, 20 (Oct. 2018).


Zheng, J., Zhang, X., Zhao, X., Tong, X., Hong, X., Xie, J., and Liu, S.
Deep-rbppred: Predicting rna binding proteins in the proteome scale based on deep learning.
Scientific reports 8, 1 (2018), 15264.

Bibliography IV

Zhou, J., and Troyanskaya, O. G.
Predicting effects of noncoding variants with deep learning–based sequence model.
Nature methods 12, 10 (2015), 931.

[39] Title is a WIP.
[40] Title is a WIP.
[41] Title is a WIP.