A maximally noncommittal physicist looks at genetic interactions

Abel Jansma, 2nd year PhD candidate
Supervised by **Chris Ponting**, **Luigi Del Debbio** & **Ava Khamseh**

BG section meeting - 05.05.20

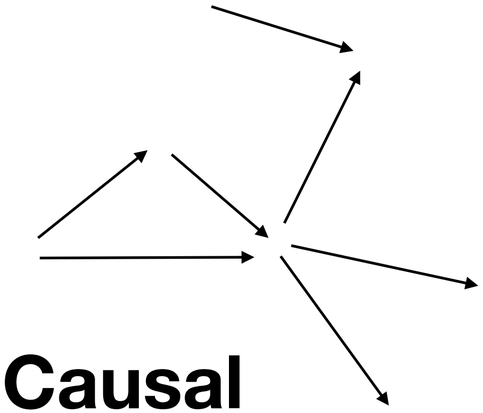
Background

Learning from genetic data

Statistical

- Pairwise correlation
- Covariance
- Higher moments
- GWAS summary stats
- **Symmetric**

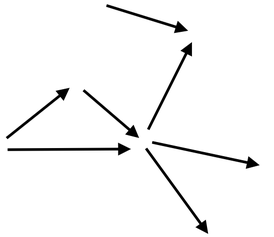
- Easy
- No assumptions
- No direct interpretation
- Cor. <=/=> Caus.
  - **In both directions**



Causal

- Average treatment effect
- Mendelian randomisation
- Graphical models, Bayes nets
- Do-calculus
- **Asymmetric/directed**

- Hard
- Assumes model
- Interpretable
- Predictive



Causal explanations

Depend on conditional independence: $p(x,\ y|z) = p(x|z)p(y|z)$

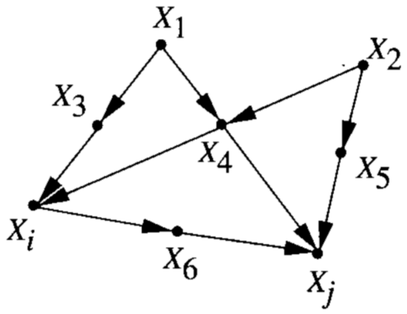


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$.

From: J. Pearl - Causality: Models, Reasoning, and Inference (2000)

Hard when

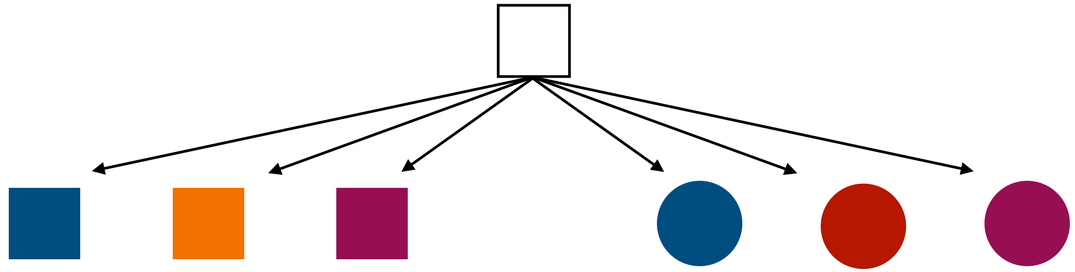
1. There are many variables
2. There is not a lot of information
3. The graph is dense
4. You're a physicist still coming to grips with the complexity of biology

Quick detour

Which will lead to something in between statistics and causality

Fairness

The principle of insufficient reason



- Experiment with 6 possible outcomes
- Assigning probabilities to outcomes?

1/6	1/6	1/6	1/6	1/6	1/6

- What if new information comes to light?
  - e.g. outcome is □

1/3	1/3	1/3	0	0	0

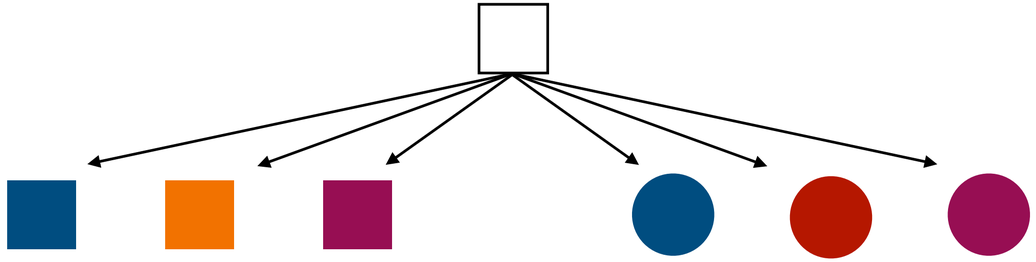
  - e.g. outcome is 

1/2	0	0	1/2	0	0

- What does fairest mean? Puzzled Laplace, Bernoulli, Boole, Venn, Keynes.

Fairness

And the case for noncommittal



- What do we do when we redistribute probabilities?
- We maximise entropy of the final distribution, given the information.
- In general, fairest possible distribution over variables x*:*
  - Maximise $S_p = \sum_{x \in \mathcal{X}} p(x)\ log\ p(x)$

- "*[The maximum entropy estimate] is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information*"

  E.T. Jaynes, *Information Theory and Statistical Mechanics* (1957)

- Are these maximum entropy (ME) distributions interesting? Let's look at some examples.

Maximum Entropy distributions

$$S_p = \sum_{x \in \mathcal{X}} p(x)\ log\ p(x)$$

Some examples

1. Supervised inference. We want to find ME distribution for categorical variables subject to constraints:

- $\sum_u p_u(x) = 1$
- $p_u(x) \geq 0\ \forall x\ \forall u$

- It predicts label u for data x
- ME distrubution: $p(x)_u = \frac{\exp(\lambda_u x)}{\sum_v \exp(\lambda_v x)}$

- i.e. logistic regression, and in turn classic feed-forward neural networks

Maximum Entropy distributions

$$S_p = \sum_{x \in \mathcal{X}} p(x)\ log\ p(x)$$

Some examples

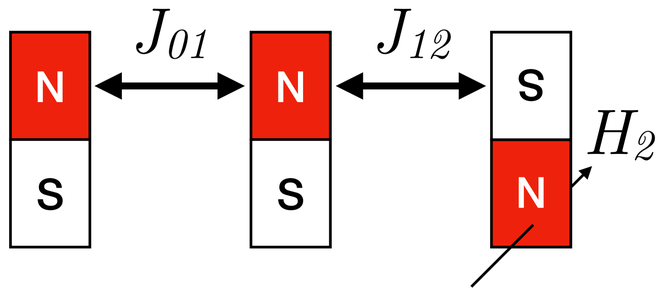
2. Unsupervised inference. We want to find ME distribution, but all we have is observations of the data, and its moments:

- $\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) = 1$
- $\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) s_i = \langle s_i \rangle_{\text{obs.}}$
- $\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) s_i s_j = \langle s_i s_j \rangle_{\text{obs.}}$

ME distribution: $p(\mathbf{s}) = \mathcal{Z}^{-1} \exp(-E(\mathbf{s}))$ where $E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j$

Maximum Entropy distributions

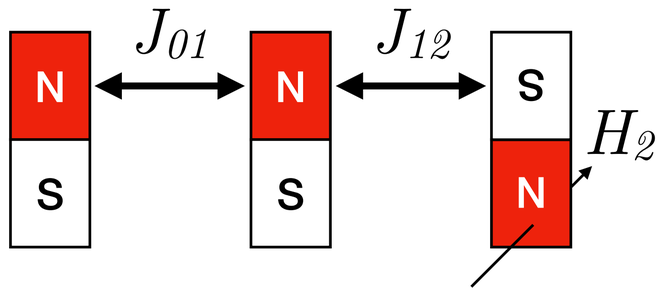
The Ising model



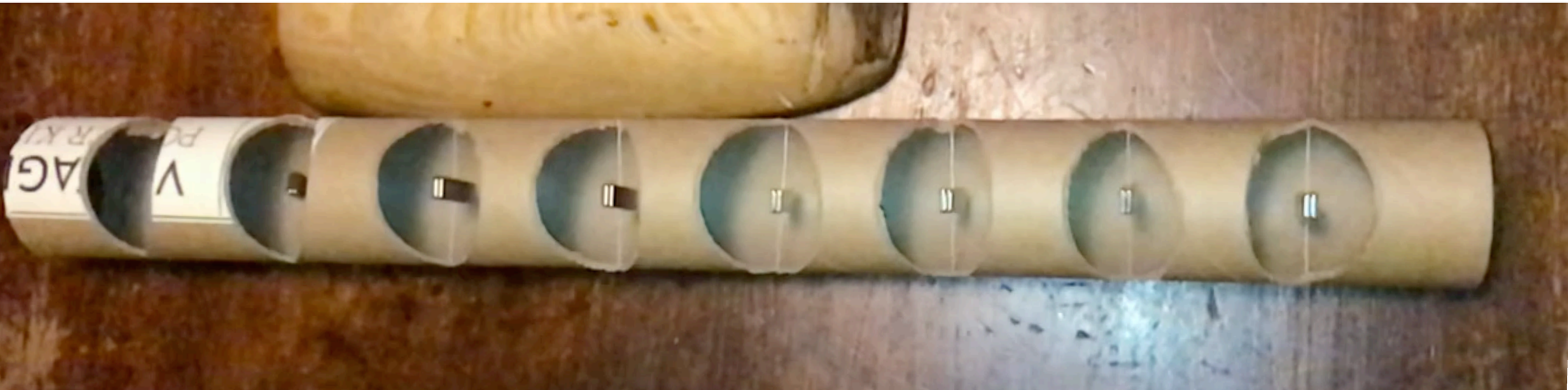
- $p(\mathbf{s}) = \mathcal{Z}^{-1} \exp(-E(\mathbf{s}))$ where $E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j$
- This is a famous model in physics, and describes ‘magnets’ in an external magnetic field H, interacting with each other through J.
- Why interactions? $E(\mathbf{s}) \propto J_{ij} s_i s_j \implies \frac{\partial E(\mathbf{s})}{\partial s_i} = J_{ij} s_j$
- Parameters $H,\ J$ can be estimated from data.
- Inverse problem used in statistics, and e.g. neuroscience.
- Something in between statistical quantity and causal…

Maximum Entropy distributions

The Ising model

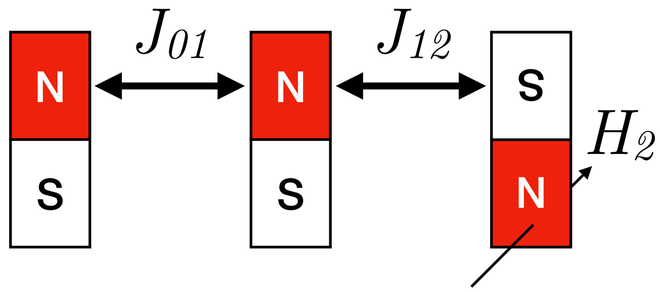


$$p(\mathbf{s}) = \mathcal{Z}^{-1} \exp(-E(\mathbf{s})) \quad \text{where } E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j$$



Maximum Entropy distributions

The generalised Ising model



- $p(\mathbf{s}) = \mathcal{Z}^{-1} \exp(-E(\mathbf{s}))$ where $E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j$
- Ising model is ME distribution after observing 2 moments.
- What about observing N moments?
- Natural extension: $E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j + \sum_{ijk} K_{ijk} s_i s_j s_k + \sum_{ijkl} L_{ijkl} s_i s_j s_k s_l + \ldots$
- Higher order interactions $K,\ L,\ \ldots$
- Already hard to estimate first two moments. Any hope for higher orders?

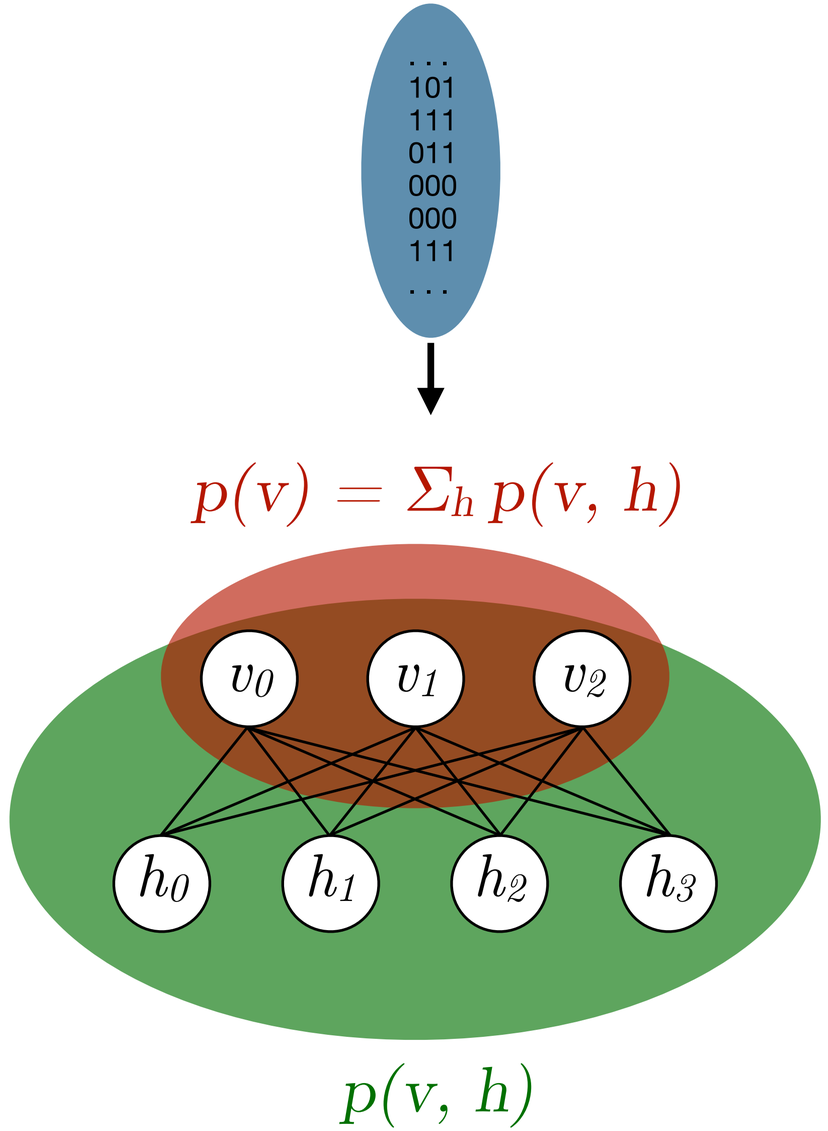
Estimating parameters

Restricted Boltzmann machines

- Two layer neural network
- Itself and Ising model

$$p(\mathbf{v}, \mathbf{h}) = \mathcal{Z}^{-1} \exp(-\sum_{ij} h_i w_{ij} v_j - \sum_j b_j v_j - \sum_i c_i h_i)$$

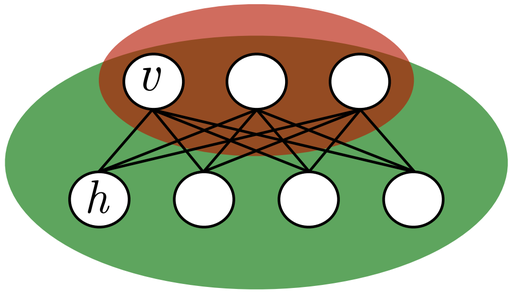
- Make marginal over one of the layers approximate data distribution
- Has to encode interactions in hidden layer



- Marginal: $E(\mathbf{v}) = -\sum_j (b_j + f_0(\sum_i w_{ij}))v_j - \sum_n \sum_{j_1 \ldots j_n} f_n(\sum_i w_{ij_1} \ldots w_{ij_n}) v_{j_1} \ldots v_{j_n}$

Estimating parameters

Restricted Boltzmann machines



- RBM: $E(\mathbf{v}) = -\sum_j (b_j + f_0(\sum_i w_{ij}))v_j - \sum_n \sum_{j_1 \ldots j_n} f_n(\sum_i w_{ij_1} \ldots w_{ij_n}) v_{j_1} \ldots v_{j_n}$
- Ising: $E(\mathbf{s}) = \sum_i H_i s_i + \sum_{i,j} J_{ij} s_i s_j + \sum_{ijk} K_{ijk} s_i s_j s_k + \sum_{ijkl} L_{ijkl} s_i s_j s_k s_l + \ldots$

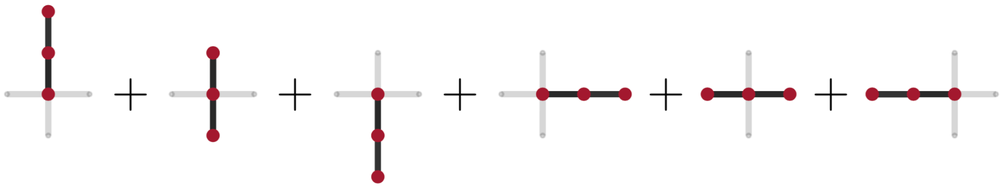
- => We can estimate higher order interactions after training RBM.
- Does this work well?

Results

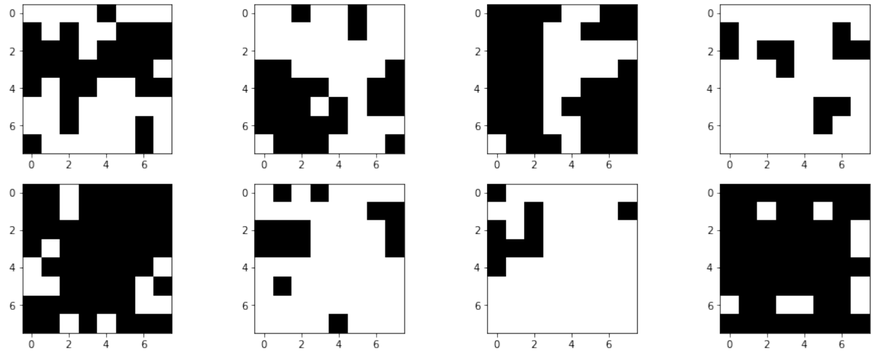
Ising model data

- Simulate interacting Ising magnets (nearest neighbours only), 100k samples
- Extract encoded interactions from RBM

- What about higher orders?
- Simulate 3-pt interactions:

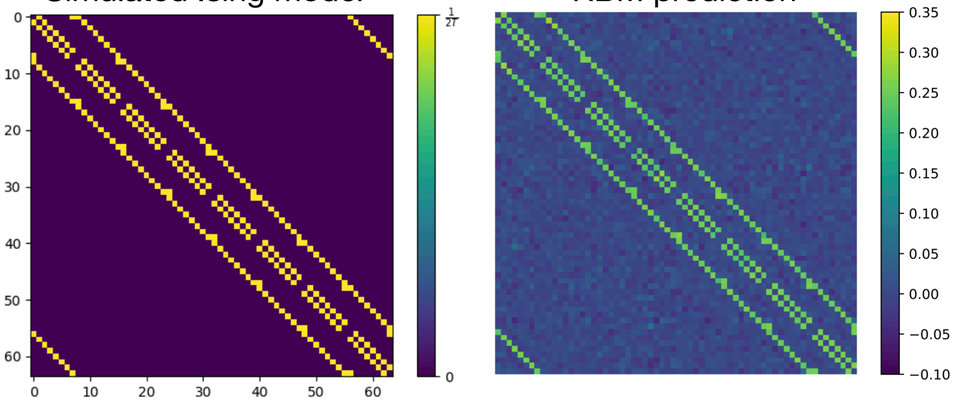


Shown: Eight sampled lattice states



Simulated Ising model

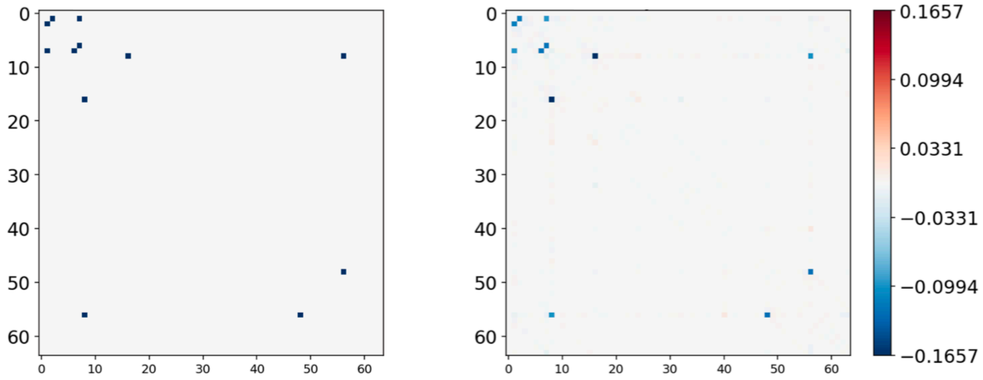
RBM prediction



From: Cossu et al. (2018)

Simulated K_{0ij}

Mean RBM prediction



Results

On to biology (finally)

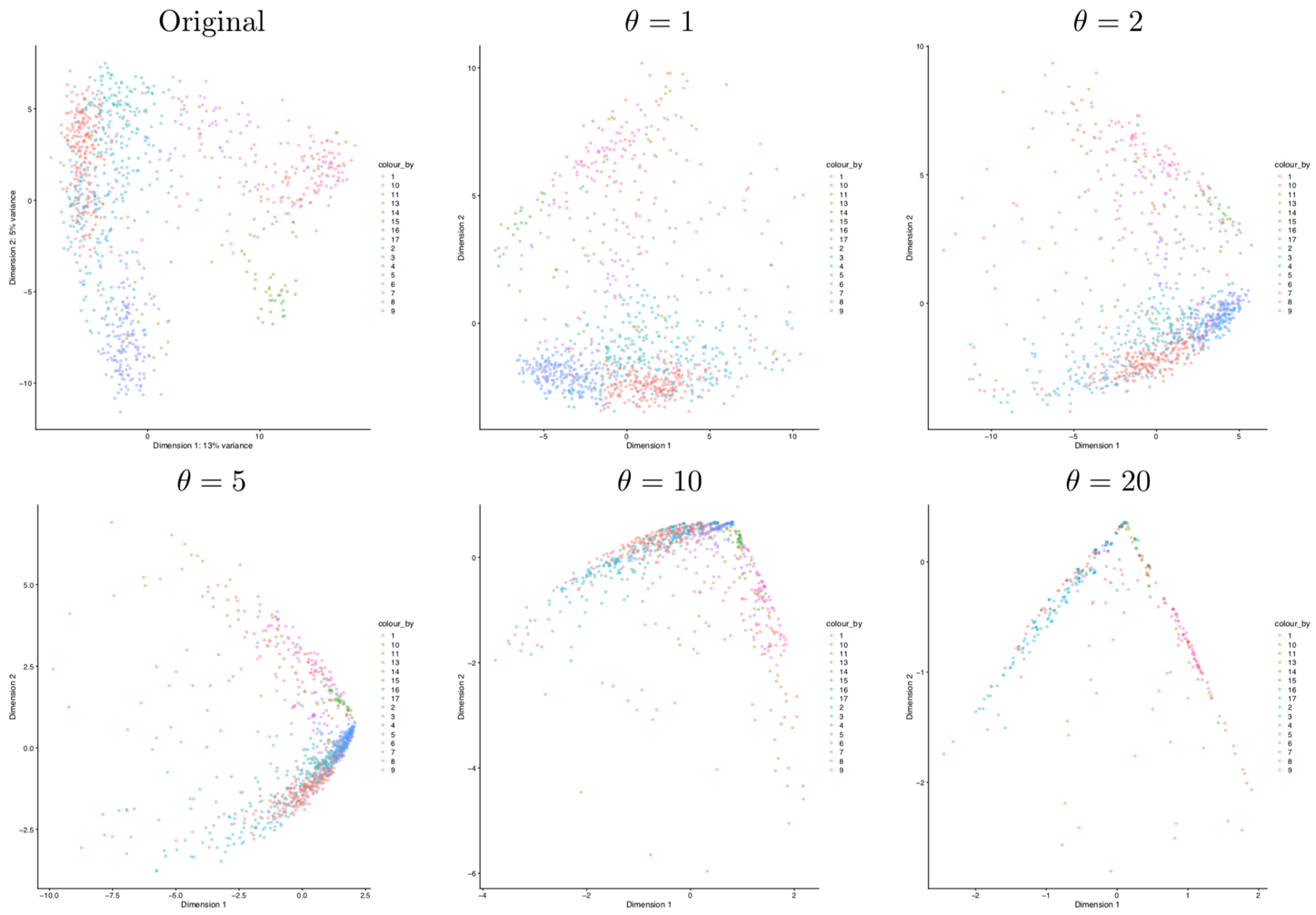
- Translate problem into genetics:
    - Magnets ⟶ Genes
    - Pole direction ⟶ Gene expression
    - Sample ⟶ Cell

- We need data on gene expression in many, many cells
    - scRNA-seq

Results

scRNA-seq data

- 10X Million Cell Data Set
- 2 mice, E18.5, 1.3M brain cells, 18.5k reads/cell
- We need to binarise data
- Put threshold at 1 count.

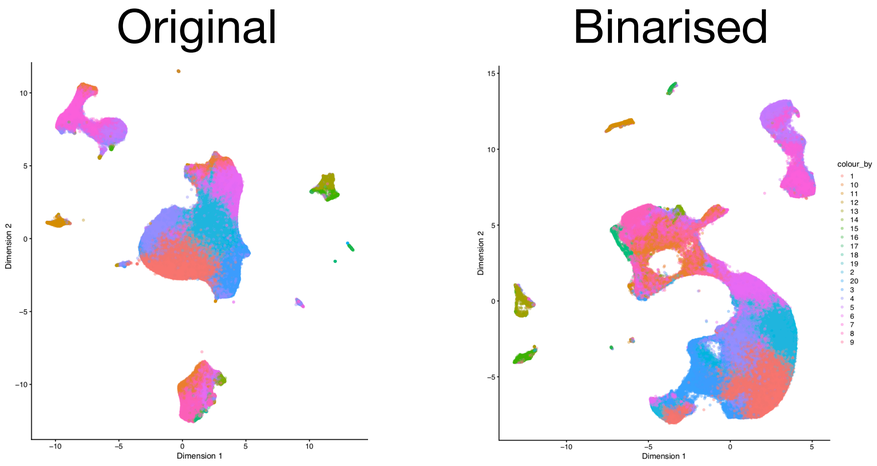


PCA embedding of different thresholds

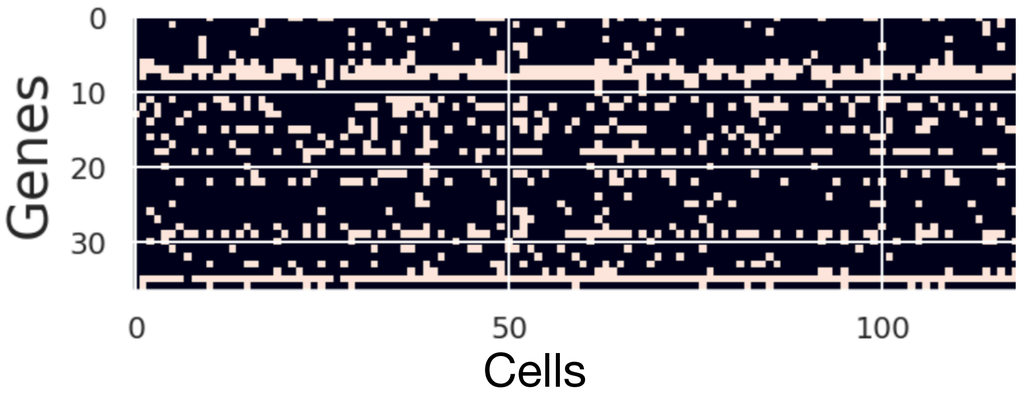
Results

scRNA-seq data

- We start with homogeneous subset:
  - ~80k astrocytes (cluster annotation)
  - We are looking for interactions, so focus on genes that play a role in astrogliogenesis
  - 37 genes (Tiwari et al., 2018, Cell Stem Cell 23, 557–571)
    - downregulated upon Atf3 & Nfia depletion
    - reduction in expression marker H3K27ac
    - showed binding sites for both Atf3 & Nfia

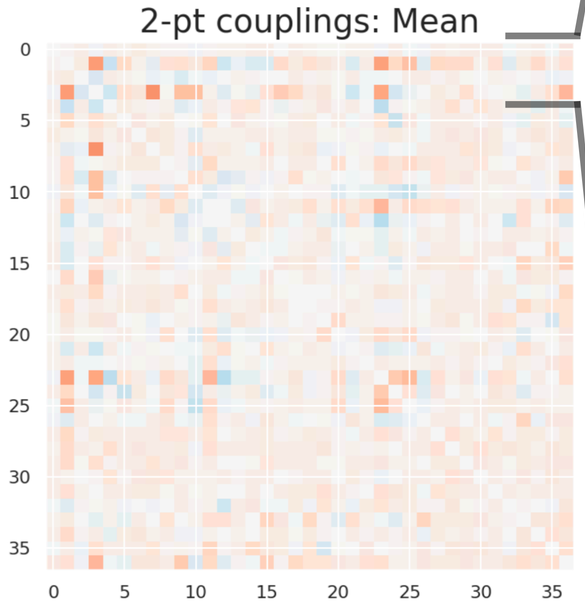


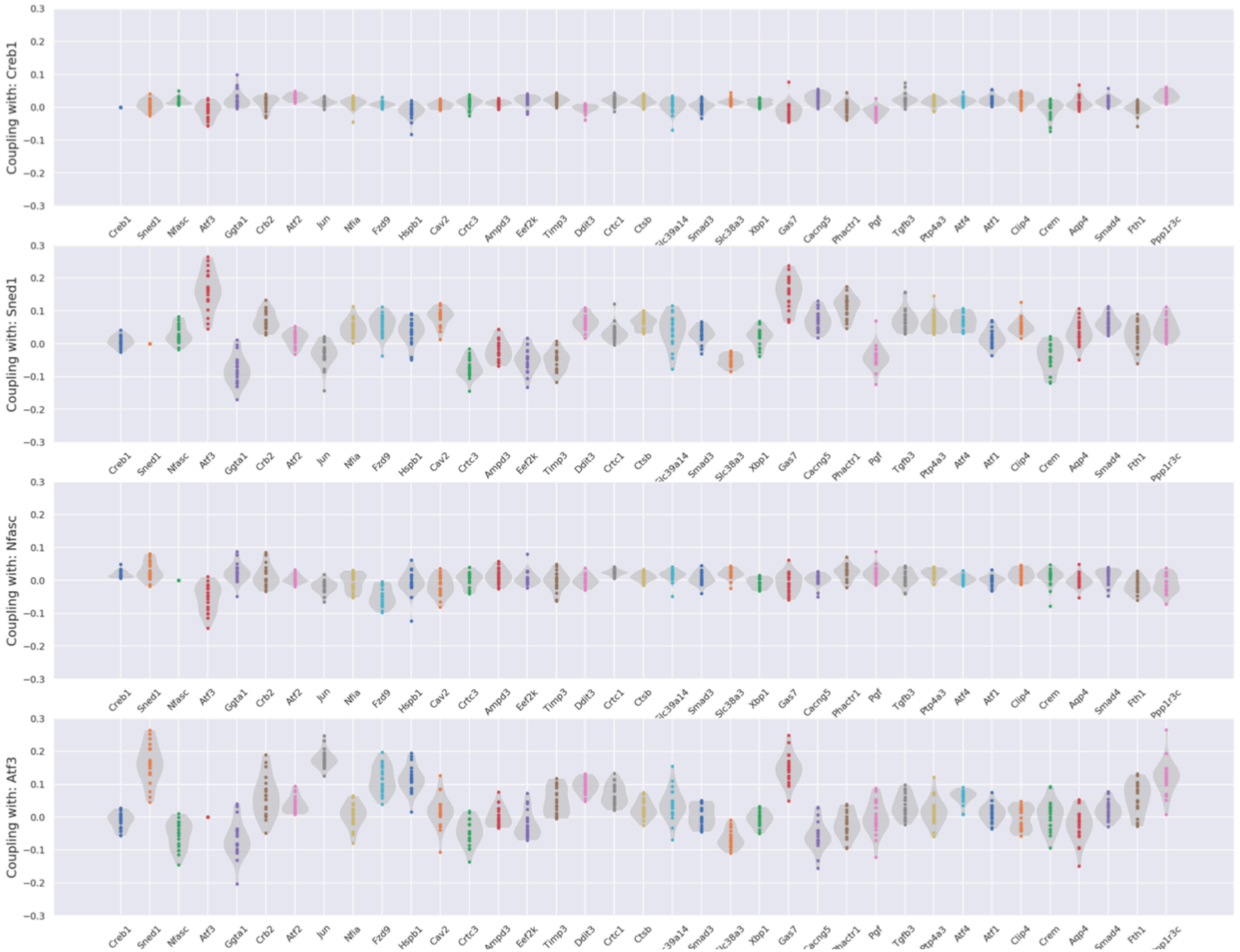
UMAP embeddings



Results

Couplings

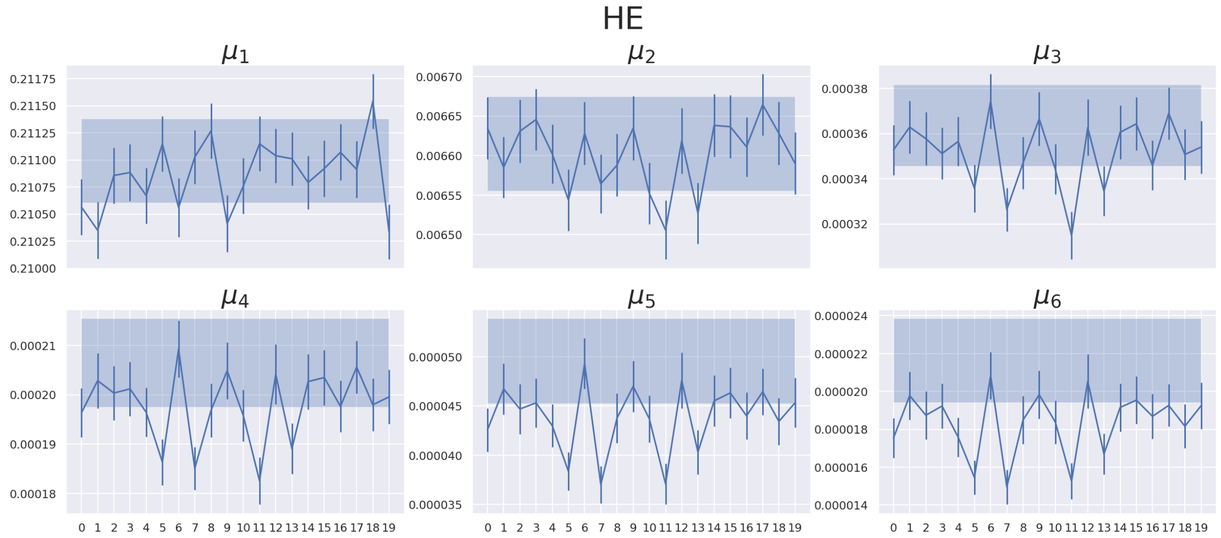


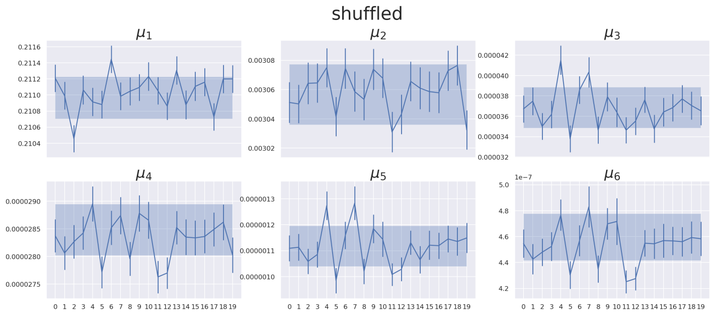


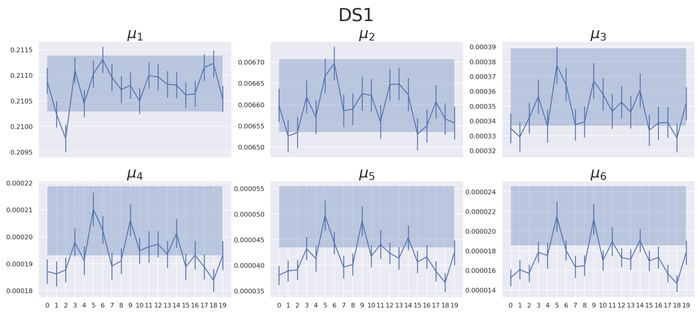
Results

Validation

- No longer have ground truth
- How well have we trained?
- Usually log-likelihood.
- We compare moments of RBM distribution to data distrution
- Full data, two disjoint subsets, and shuffled data



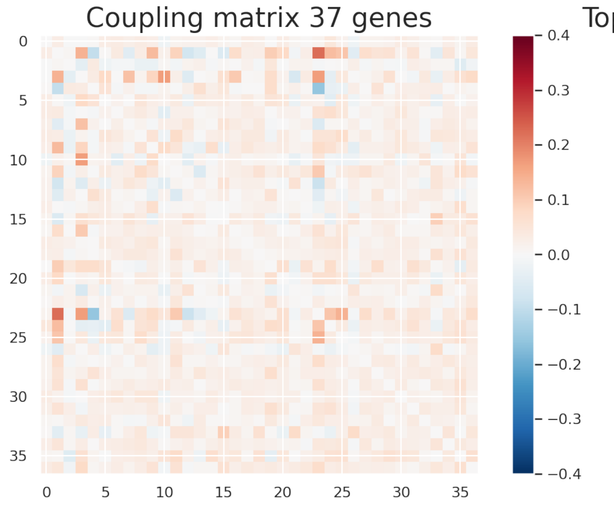


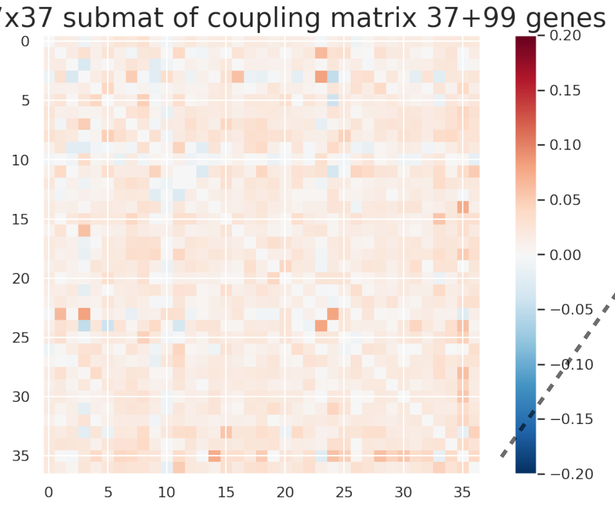


Results

Robustness

- Adding top 100 HVGs
- Original in top left corner

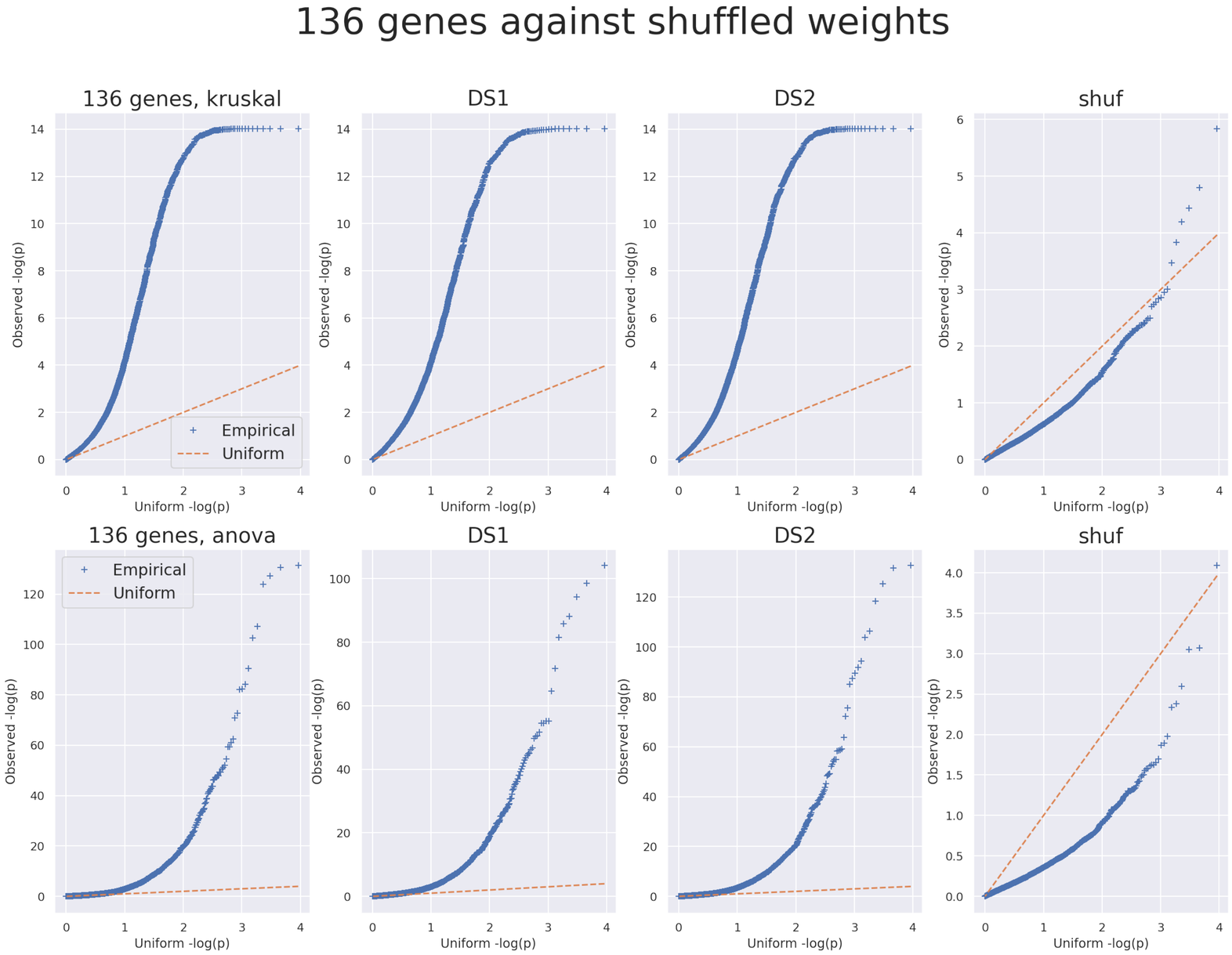




Results

Significance

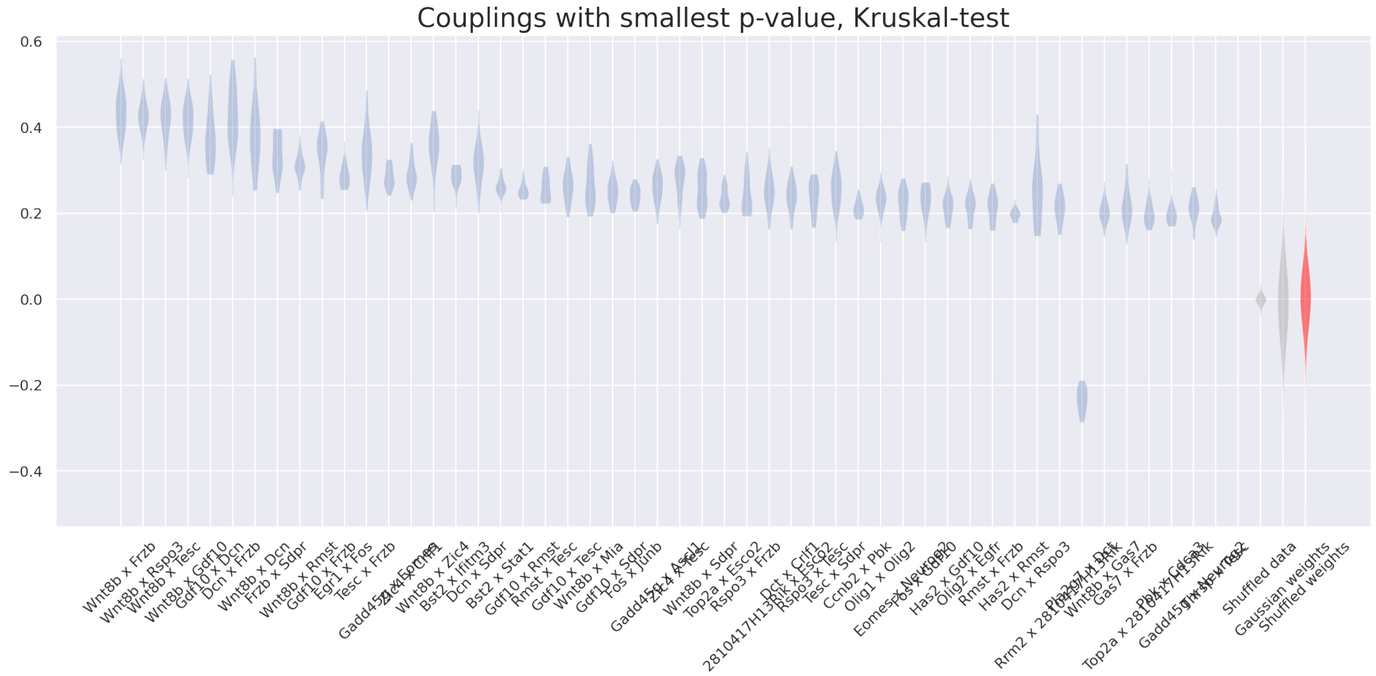
- Can assign p-value to each coupling
- QQ-plots are odd
- Couplings not independent random variables

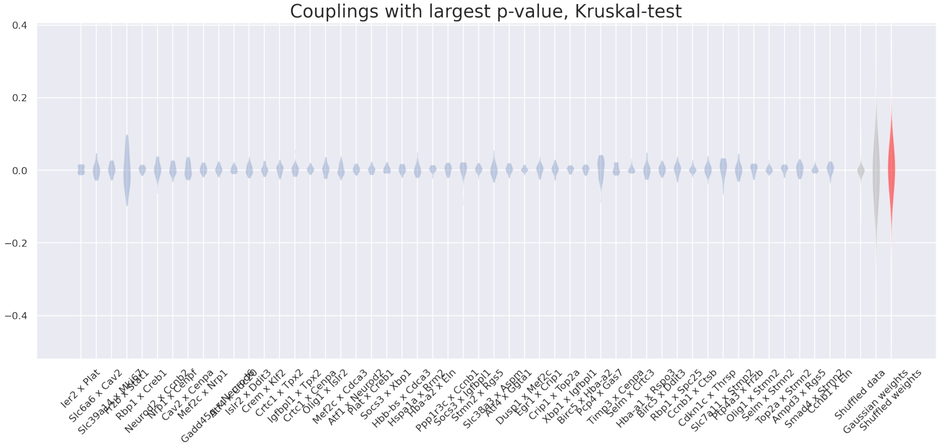


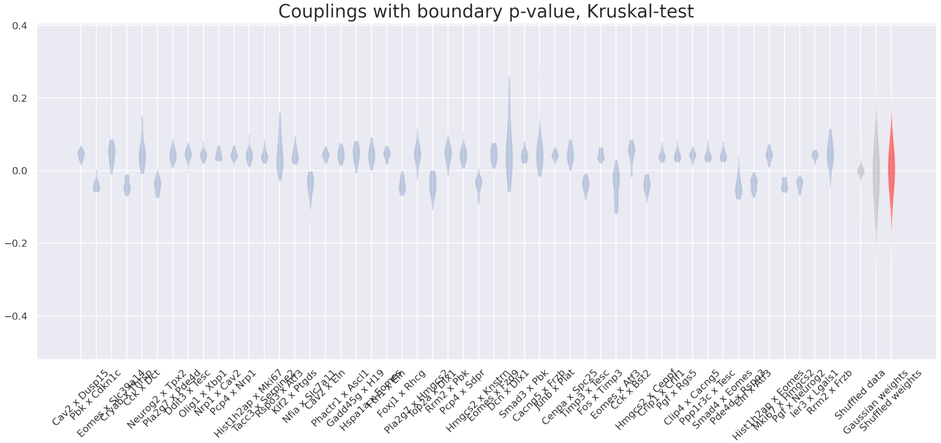
Results

Significance

- Still, rankings seem reasonable







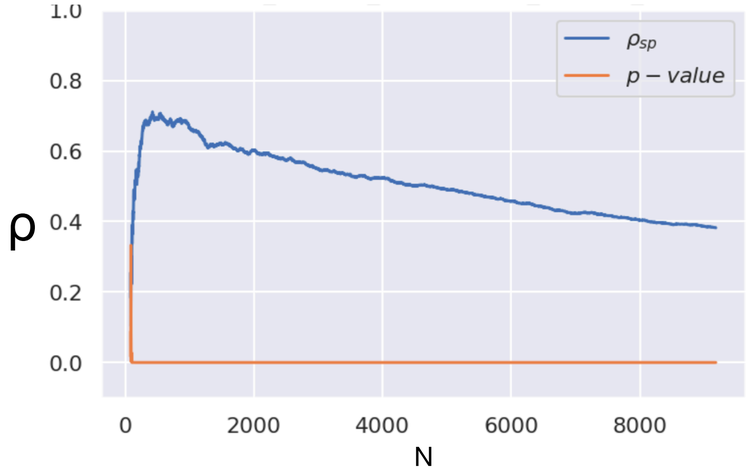
Results

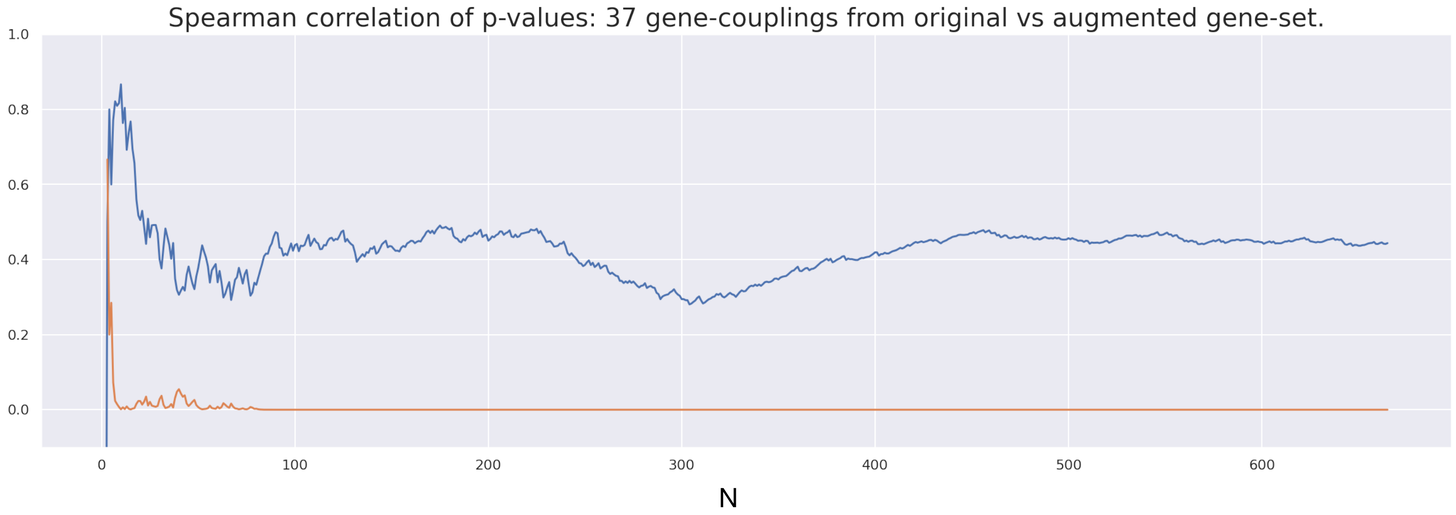
Significance

- We also want these rankings to be consistent.
- Let’s look at spearman correlation of the first N couplings, as function of N:

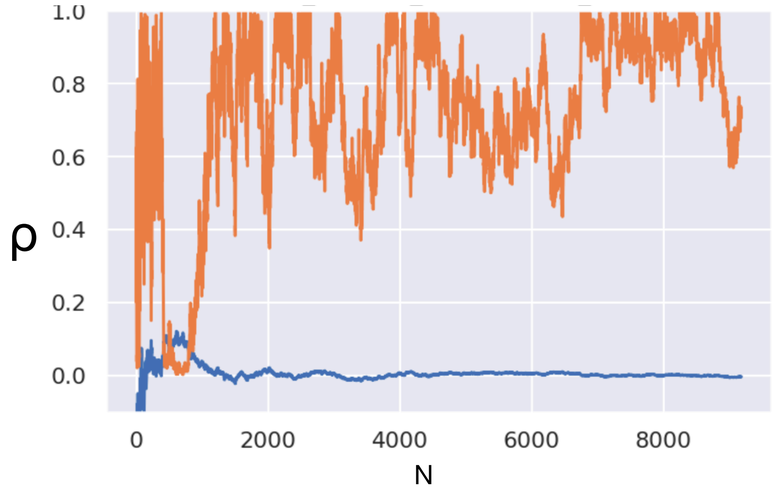
Spearman Correlation ρ

DS1 v DS2





Shuffled v orig.



Next steps

A look ahead

- I need a systematic way to select significant couplings
    - Significance threshold doesn’t work, FDR doesn’t work, what will?
- Then I can start comparing pairwise interactions to triplets etc.
- Is there a way to validate predictions? (STRING database etc.)
- **Other Cell types**
    - How do networks differ?
    - Transitions between types
    - Cell atlas

Acknowledgements

Many people to thank




- Higgs Centre @KB: Luigi Del Debbio, Guido Cossu, Tommaso Giani, Michael Wilson

- Ponting lab @IGMM:

Missing from photo:
Louise Docherty
Joshua Dibble
Cath Heath