THE UNIVERSITY of EDINBURGH

Higher-order interactions in statistical physics, machine learning and biomedicine

Ava Khamseh

Biomedical AI Lab
January 2023

Collaborators

Institute of Genetics and Cancer

Abel Jansma
Yuelin Yao
Jareth Wolf
Ava Khamseh (and **School of Informatics**)
Chris Ponting

School of Physics

Luigi Del Debbio

School of Mathematics

Sjoerd Beentjes

A cross-disciplinary journey (2017-2023)

Study of interactions:

Part 1: Using a neural network to estimate interactions in Ising model

Part 2: model-independent estimation of interactions directly from data

Part 3: Biological interpretation of interactions and application in biomedicine

Forward vs inverse problem

Forward problem (Statistical Physics): The goal is to provide a macroscopic description of Nature by deriving observable quantities from underlying laws.

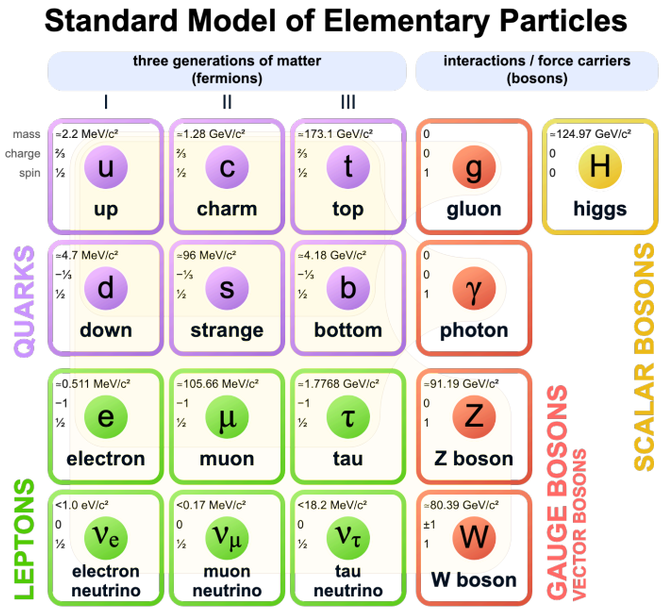
- Ising model forward problem: Obtain observables such as magnetisation, energy and correlations, given the Hamiltonian and its parameters

Inverse problem: Starting point are observations (data), the goal is to infer microscopic properties of the system

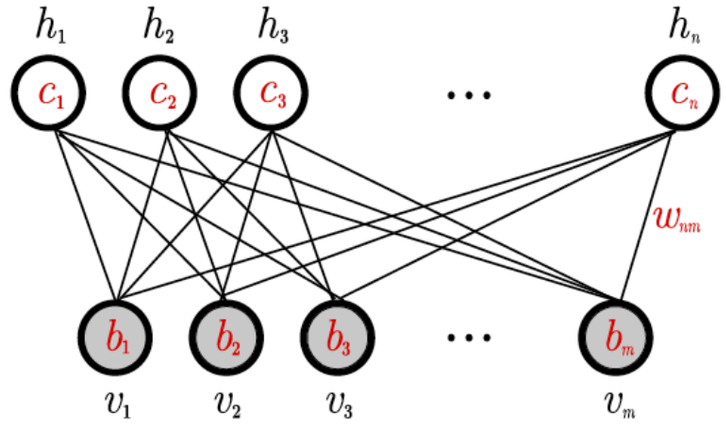
- Estimate Ising interactions directly from data

Interactions

Elementary Particles

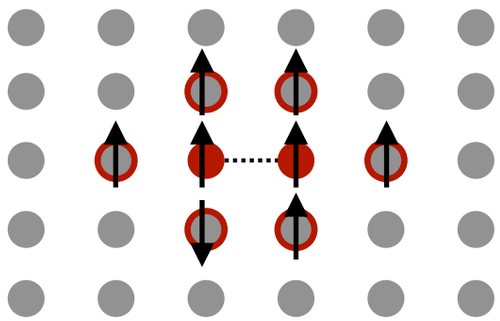


Nodes in a neural network

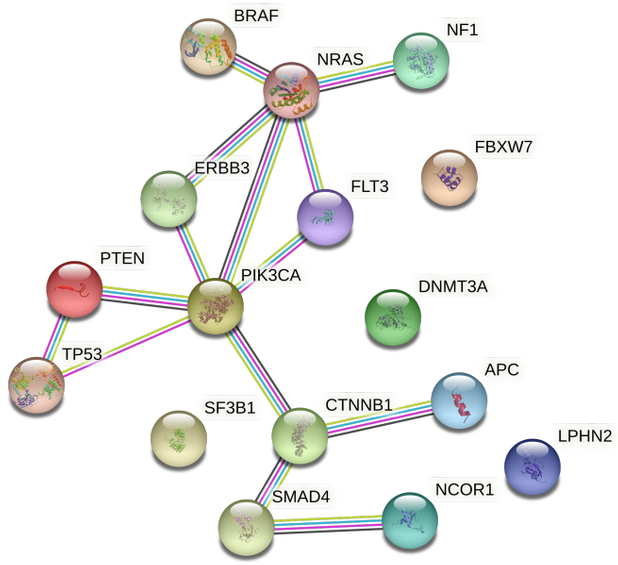


Dynamical/physical,
vs,
Probabilistic

Statistical Physics



Biomedicine



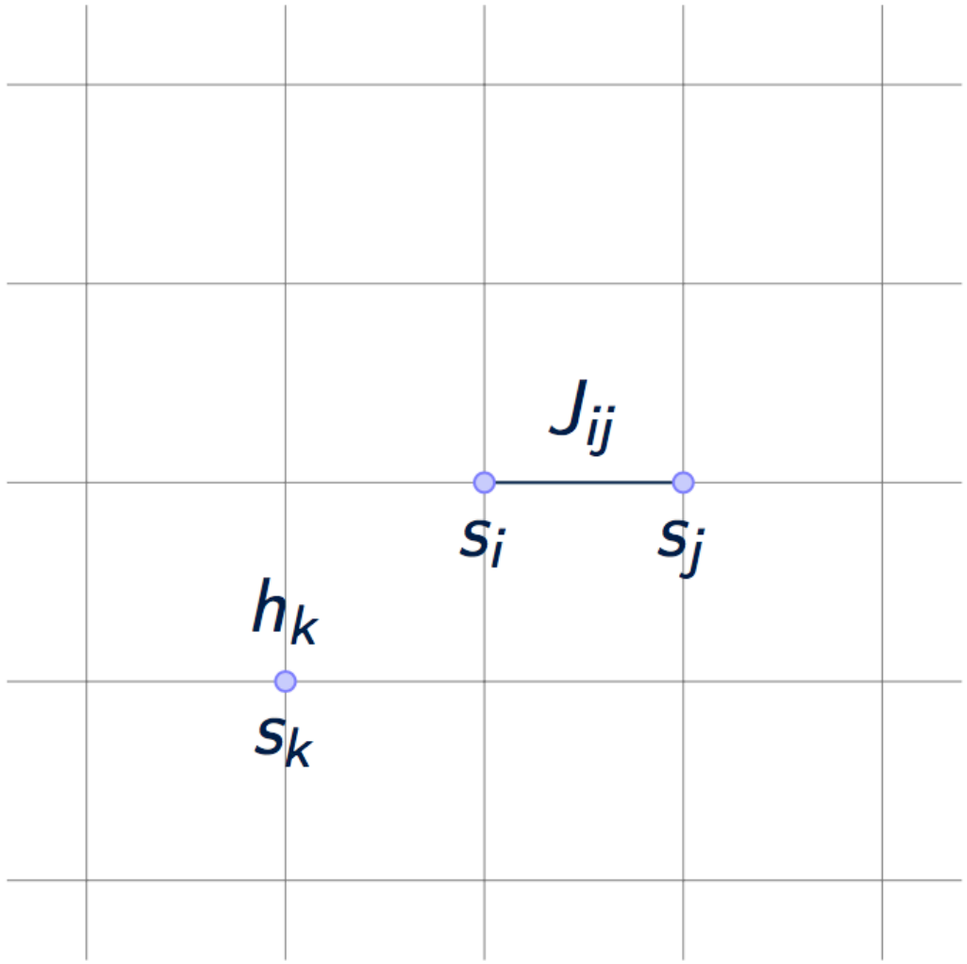
Part 1

Interactions: The Ising model & RBM

Ising Model

$$p_D(s) = \frac{1}{Z(J,h)}\, e^{-H_{J,h}(s)}$$

$$H_{J,h} = -\sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i \quad , \quad Z(J,h) = \sum_s e^{-H_{J,h}(s)}$$



MC simulation of the 2D Ising model at various temperatures

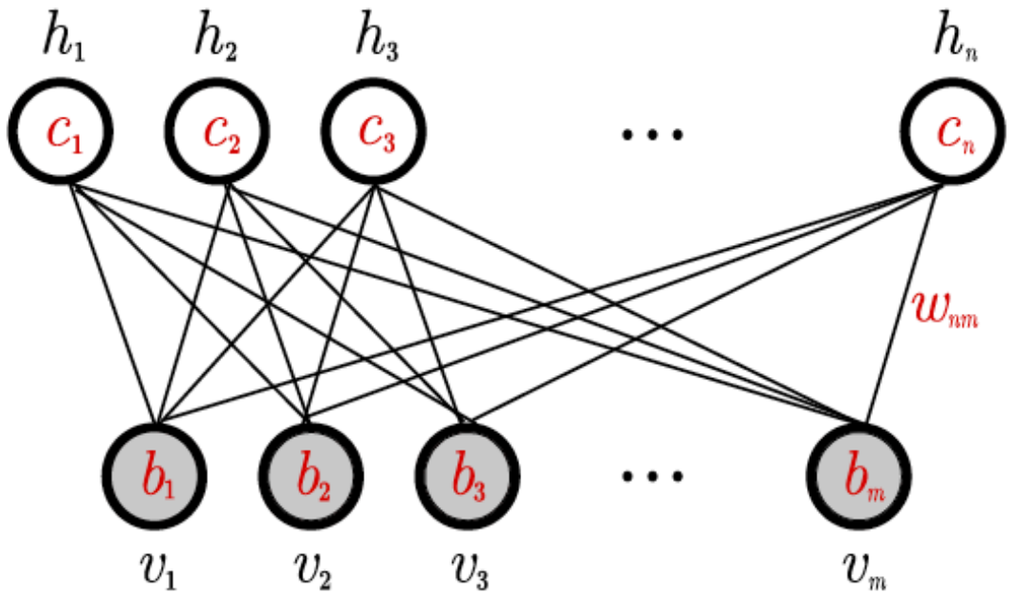
generate a sample
$D = \{s^1, s^2, \ldots\}$, $N_D \sim 10^5$

Onsager 1944

Restricted Boltzmann Machine (RBM)

$$E_\theta(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{i=1}^{n} c_i h_i - \sum_{j=1}^{m} b_j v_j$$

$$\boxed{p_{\mathrm{RBM}}(\mathbf{v},\mathbf{h}|\theta) = \frac{1}{Z_{\mathrm{RBM}}} e^{-E_\theta(\mathbf{v},\mathbf{h})}}$$



Ising configuration training data

$$D_{\mathrm{KL}}\left(q_{\mathrm{data}}(\mathbf{v})||p_{\mathrm{RBM}}(\mathbf{v}|\theta)\right) = \sum_{\mathbf{v}} q_{\mathrm{data}}(\mathbf{v}) \log\left(\frac{q_{\mathrm{data}}(\mathbf{v})}{p_{\mathrm{RBM}}(\mathbf{v}|\theta)}\right)$$

$$= \sum_{\mathbf{v}} \left(q_{\mathrm{data}}(\mathbf{v}) \log\left(q_{\mathrm{data}}\right) - q_{\mathrm{data}}(\mathbf{v}) \log\left(p_{\mathrm{RBM}}(\mathbf{v}|\theta)\right) \right)$$

Max likelihood $\iff$ Min KL divergence

Observables

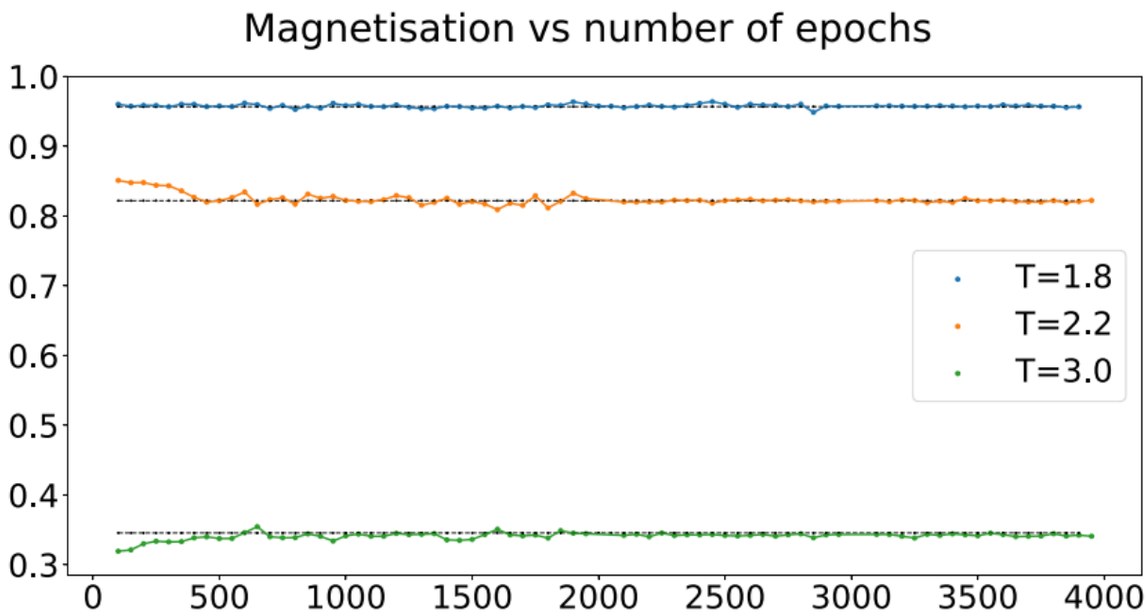
$$\langle m \rangle = \frac{1}{L^2} \left\langle \left| \sum_{i=1}^{L^2} s_i \right| \right\rangle ,$$

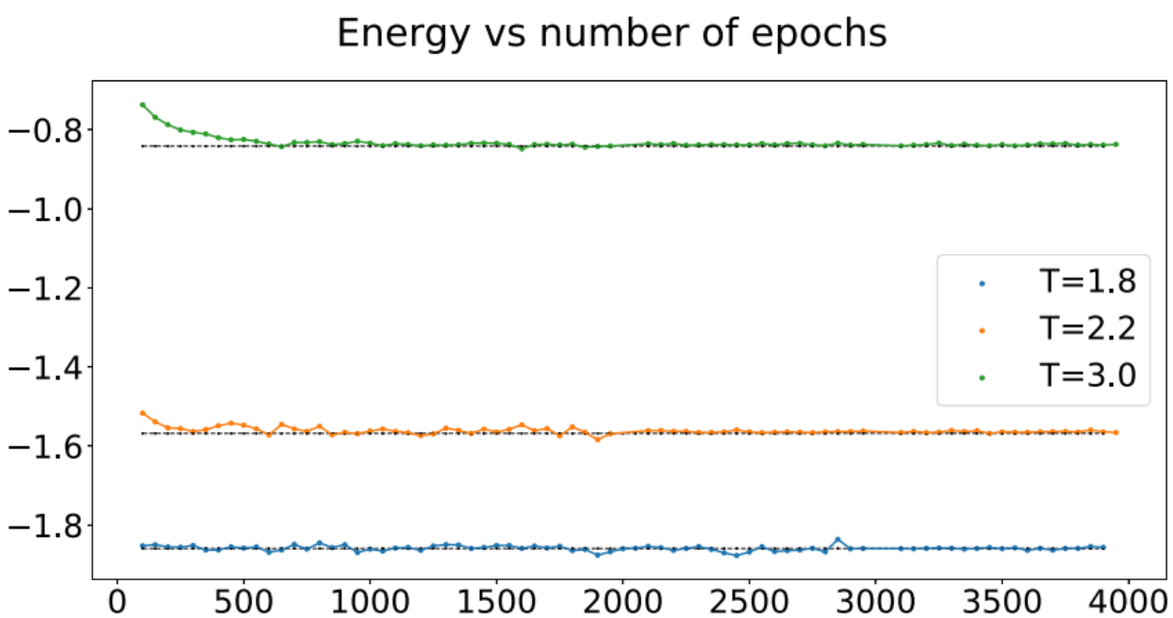
$$\langle \chi \rangle = \frac{L^2}{T} \left\langle \langle m^2 \rangle - \langle m \rangle^2 \right\rangle ,$$

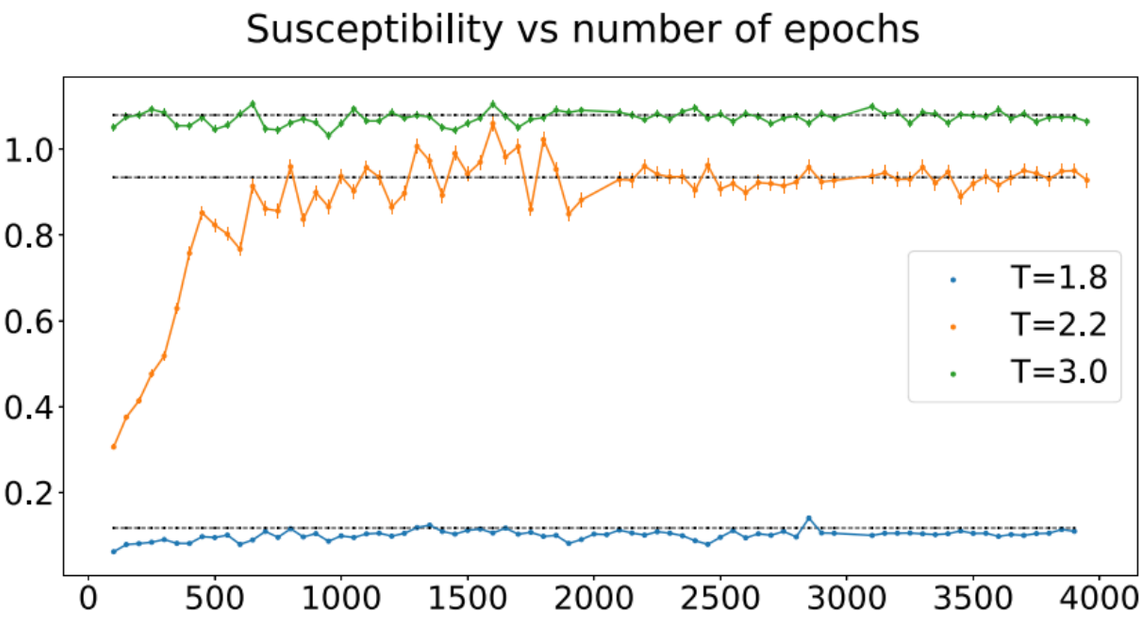
$$\langle E \rangle = -\frac{1}{L^2} \left\langle \sum_{\langle i,j \rangle} s_i s_j \right\rangle ,$$

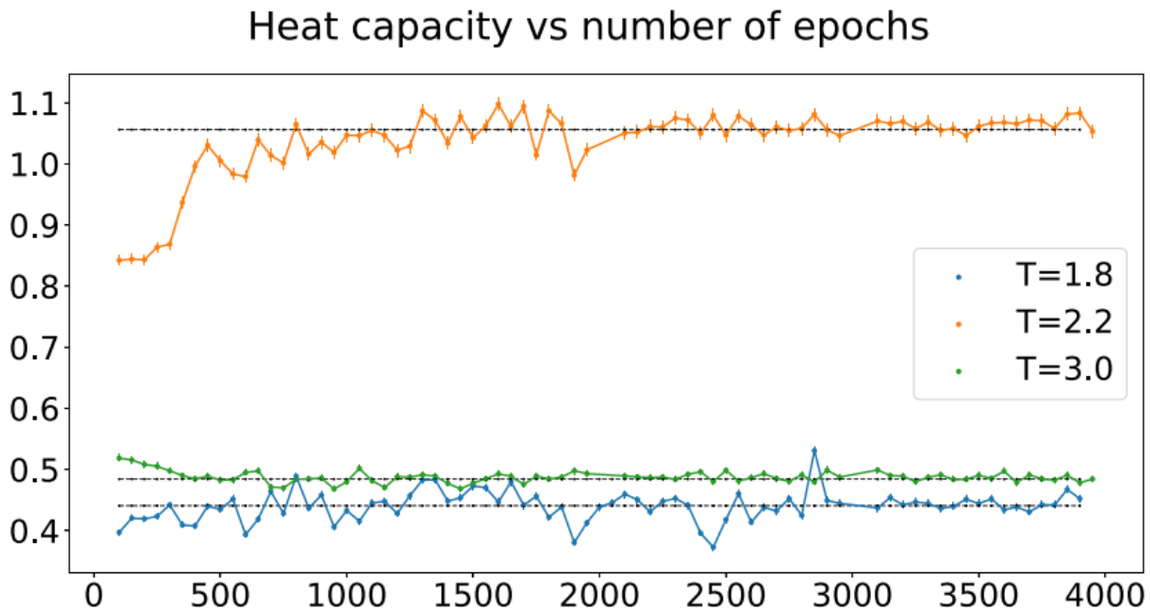
$$\langle c_v \rangle = \frac{L^2}{T^2} \left\langle \langle E^2 \rangle - \langle E \rangle^2 \right\rangle .$$

RBM: Observables



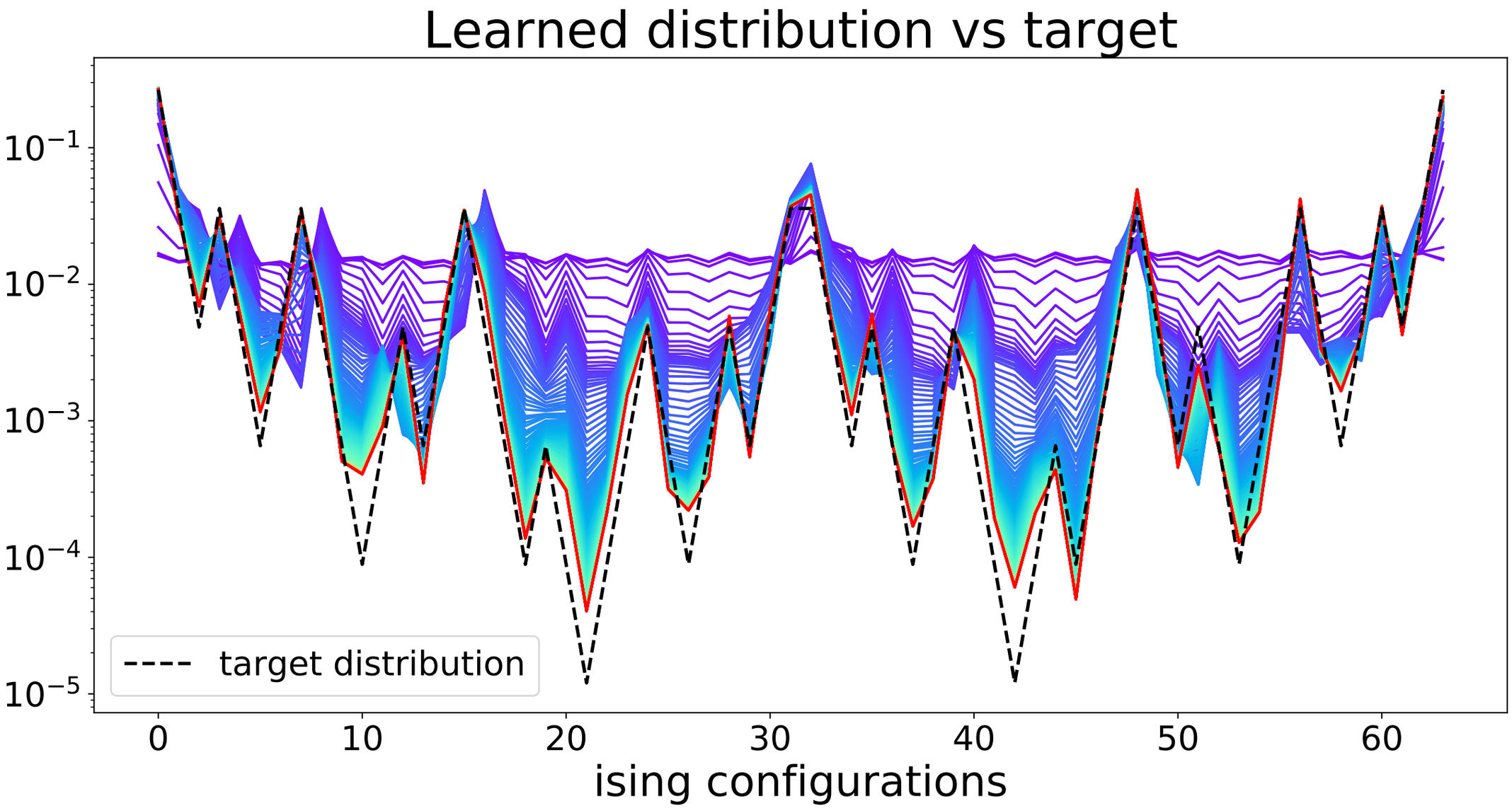






Cossu et. al., Physical Review B (2018)

Example: 1D Ising model in 6 variables



RBM: Estimation of interactions

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_j \left(\sum_i \kappa_i^{(1)} W_{ij} \right) v_j - \frac{1}{2} \sum_{jk} \left(\sum_i \kappa_i^{(2)} W_{ik} W_{ij} \right) v_j v_k + \cdots$$

Beyond pairwise, higher-order couplings

Possible to re-sum the entire series to obtain 2-point coupling!!

For binary data, using cumulant generating function …

$$v_j^n = v_j \quad , \quad n \in \mathbb{Z}^+$$

finite sum

RBM: Estimation of interactions

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_j \left(\sum_i \kappa_i^{(1)} W_{ij} \right) v_j - \frac{1}{2} \sum_{jk} \left(\sum_i \kappa_i^{(2)} W_{ik} W_{ij} \right) v_j v_k + \cdots$$

Beyond pairwise, higher-order couplings

Possible to re-sum the entire series to obtain 2-point coupling!!

For binary data, using cumulant generating function …

$$v_j^n = v_j \quad , \quad n \in \mathbb{Z}^+$$

finite sum

e.g., 2-point interaction:

$$H_{j_1 j_2} = \frac{1}{8} \sum_i \ln \frac{(1 + e^{c_i + W_{ij_1} + W_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + W_{ij_1}})(1 + e^{c_i + W_{ij_2}})}$$

Closed form expression!

RBM: Estimation of interactions

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_j \left(\sum_i \kappa_i^{(1)} W_{ij} \right) v_j - \frac{1}{2} \sum_{jk} \left(\sum_i \kappa_i^{(2)} W_{ik} W_{ij} \right) v_j v_k + \cdots$$

Beyond pairwise, higher-order couplings

Possible to re-sum the entire series to obtain 2-point coupling!!

For binary data, using cumulant generating function …

$$v_j^n = v_j \quad , \quad n \in \mathbb{Z}^+$$

finite sum

e.g., 3-point interaction:

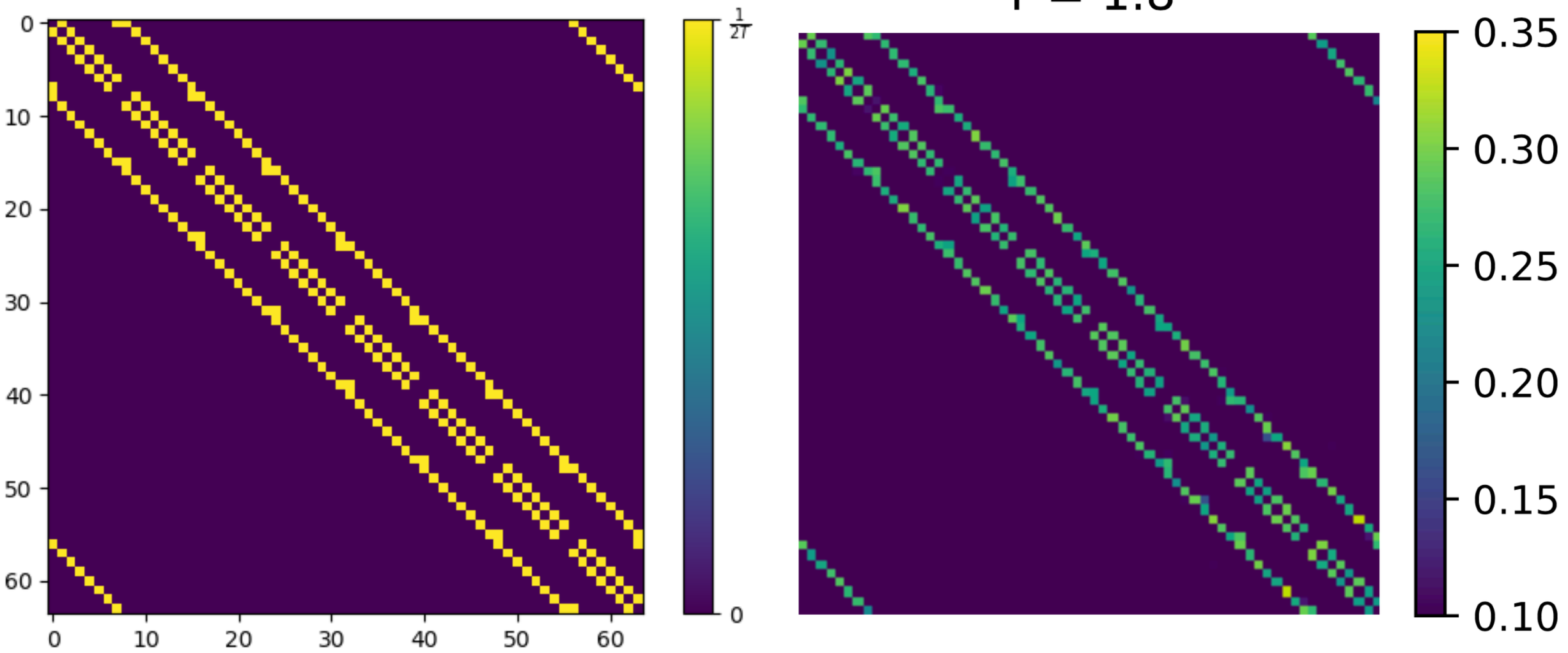
$$\frac{1}{6} \sum_i \ln \frac{(1+e^{c_i+W_{ij_1}+W_{ij_2}+W_{ij_3}})(1+e^{c_i+W_{ij_1}})(1+e^{c_i+W_{ij_2}})(1+e^{c_i+W_{ij_3}})}{(1+e^{c_i+W_{ij_1}+W_{ij_2}+})(1+e^{c_i+W_{ij_1}+W_{ij_3}})(1+e^{c_i+W_{ij_2}+W_{ij_3}})(1+e^{c_i})}$$

Closed form expression!

RBM: Couplings J_{ij}

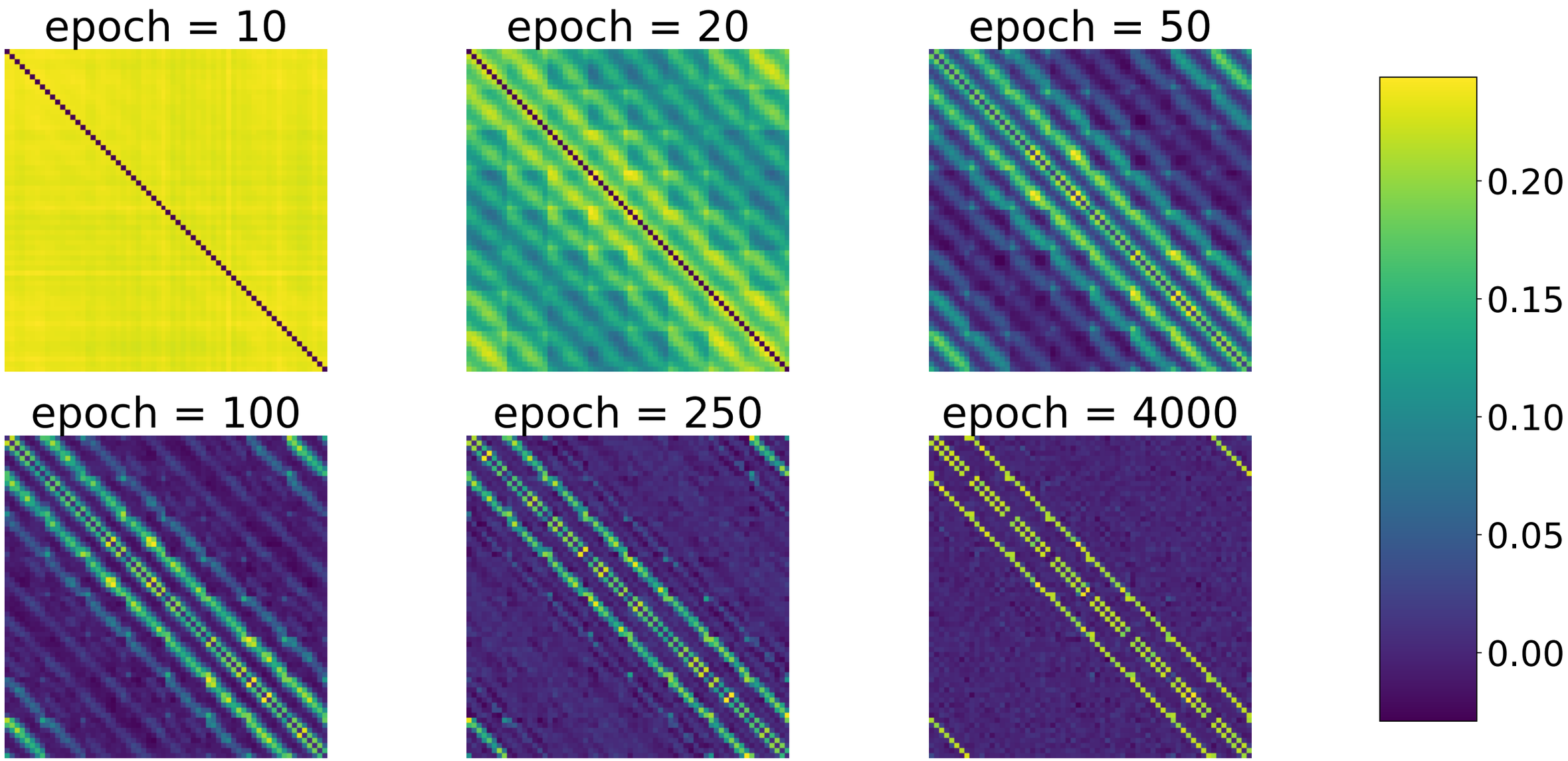
Ising Model

RBM prediction



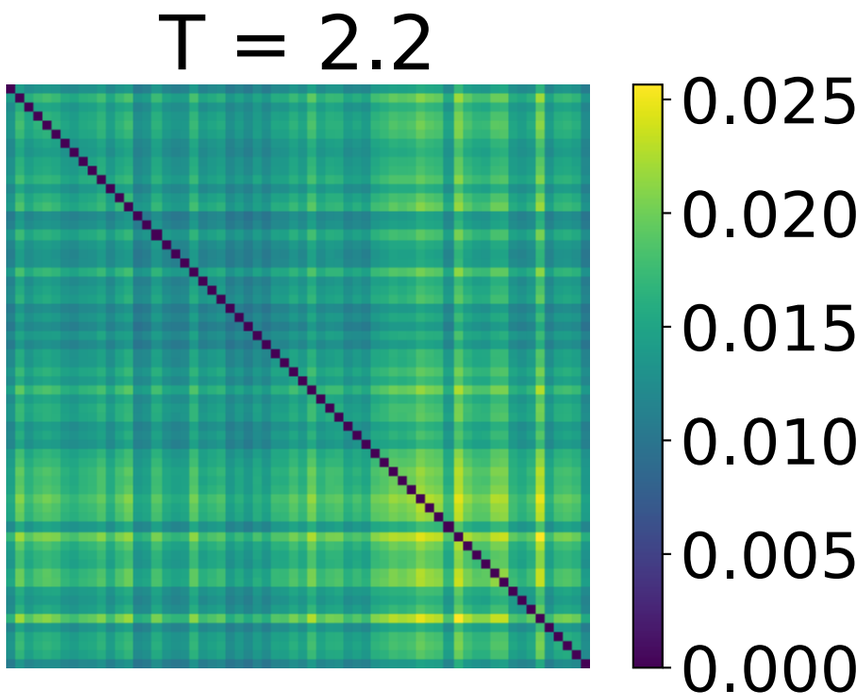
Cossu et. al., Physical Review B (2018)

RBM: Couplings during training

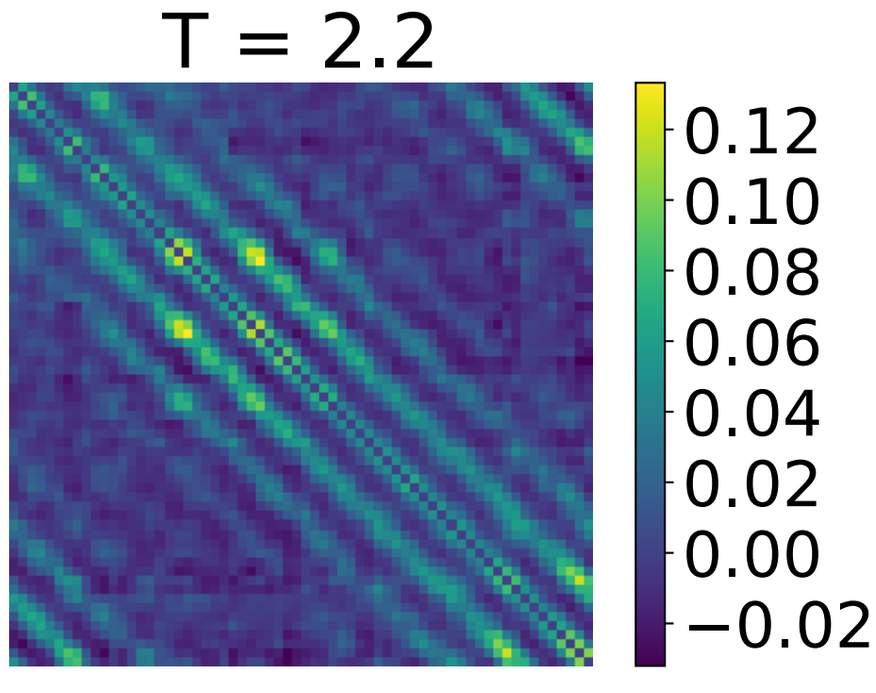


Cossu et. al., Physical Review B (2018)

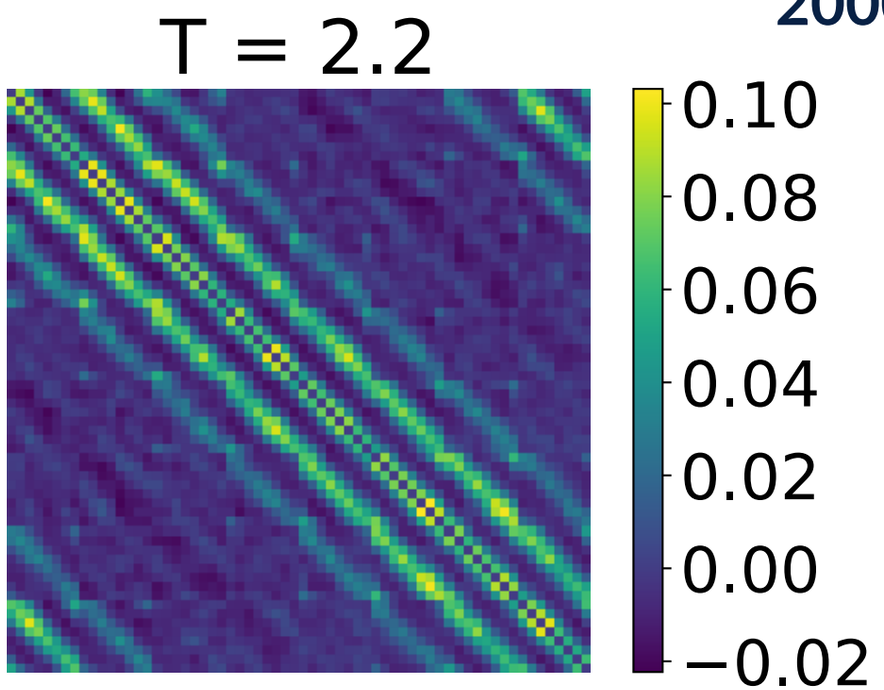
RBM: Number of training examples



200 Examples



2000 Examples



10000 Examples

Cossu et. al., Physical Review B (2018)

RBM: Lessons Learnt

- Understand well the training criteria from RBMs: Log-likelihood, Loss, free energy, reconstruction error + moments generated by the machine

- RBMs are successful at estimating (higher-order) interactions in a given system of binary variables

- Generally, need lots of training examples

AND

- Still need to deal with potentially very large numbers of dependent variables (e.g. Gene Networks). **RBM interactions changing depending on gene included**!
- RBMs are not particularly convenient to train … (e.g. including time on hyper-parameter tuning)

Part 2

Interactions: Model-independent definition and estimation

Defining the target

Aim: Formulate the **target** quantity of interest:

> **not** as a property of a parametric statistical model

The target quantity can often be identified **without** ever specifying the functional or distributional form of the model: **model-independent**

Why is this important?

Targeted Learning

1) Be clear about what we are actually after.
2) Don’t waste computational, analytical and data resources on irrelevant aspects of a problem (here the full joint distribution!)

Judea Pearl (2010), Mark van der Laan (2011)

There is no “theory” in biology

Come up with a ‘sensible’ model-independent statistical definition

For 1 spin:

$$I_i^m = \ln\left(\frac{p(G_i = 1 \mid \underline{G} = 0)}{p(G_i = 0 \mid \underline{G} = 0)}\right)$$

Here G = spins

Later, G = genes

‘odds ratio’: What is the likelihood of spin i being 1 vs 0

There is no “theory” in biology

Come up with a ‘sensible’ model-independent statistical definition

For 1 spin:

$$I_i^m = \ln\left(\frac{p(G_i = 1 \mid \underline{G} = 0)}{p(G_i = 0 \mid \underline{G} = 0)}\right)$$

Here G = spins

Later, G = genes

‘odds ratio’: What is the likelihood of spin i being 1 vs 0

For 2 spins:

$$I_{i,j}^m = \ln\left(\frac{p(G_{ij} = (1,1) \mid \underline{G} = 0)}{p(G_{ij} = (0,1) \mid \underline{G} = 0)}\right) - \ln\left(\frac{p(G_{ij} = (1,0) \mid \underline{G} = 0)}{p(G_{ij} = (0,0) \mid \underline{G} = 0)}\right)$$

‘odds ratio’ of spin i with spin j being 1

‘odds ratio’ of spin i with spin j being 0

‘generalised odds ratio’: Does the likelihood of spin i being 1 increase/decrease depending on whether spin j is 1/0. Generalisable to higher-orders.

There is no “theory” in biology

Come up with a ‘sensible’ model-independent statistical definition

For 1 spin:

$$I_i^m = \ln\left(\frac{p(G_i = 1 \mid \underline{G} = 0)}{p(G_i = 0 \mid \underline{G} = 0)}\right)$$

Here G = spins

Later, G = genes

‘odds ratio’: What is the likelihood of spin i being 1 vs 0

For 2 spins:

$$I_{i,j}^m = \ln\left(\frac{p(G_{ij} = (1,1) \mid \underline{G} = 0)}{p(G_{ij} = (0,1) \mid \underline{G} = 0)}\right) - \ln\left(\frac{p(G_{ij} = (1,0) \mid \underline{G} = 0)}{p(G_{ij} = (0,0) \mid \underline{G} = 0)}\right)$$

If two spins are independent: $p(G_i, G_j|\underline{G} = 0) = p(G_i|\underline{G} = 0)p(G_j|\underline{G} = 0)$

There is no interaction: $I_{i,j}^m = 0$

There is no “theory” in biology

Come up with a ‘sensible’ model-independent statistical definition

For 1 spin:

$$I_i^m = \ln\left(\frac{p(G_i = 1 \mid \underline{G} = 0)}{p(G_i = 0 \mid \underline{G} = 0)}\right)$$

Here G = spins

Later, G = genes

‘odds ratio’: What is the likelihood of spin i being 1 vs 0

For 2 spins:

$$I_{i,j}^m = \ln\left(\frac{p(G_{ij} = (1,1) \mid \underline{G} = 0)}{p(G_{ij} = (0,1) \mid \underline{G} = 0)}\right) - \ln\left(\frac{p(G_{ij} = (1,0) \mid \underline{G} = 0)}{p(G_{ij} = (0,0) \mid \underline{G} = 0)}\right)$$

If two spins are independent: $p(G_i, G_j|\underline{G} = 0) = p(G_i|\underline{G} = 0)p(G_j|\underline{G} = 0)$

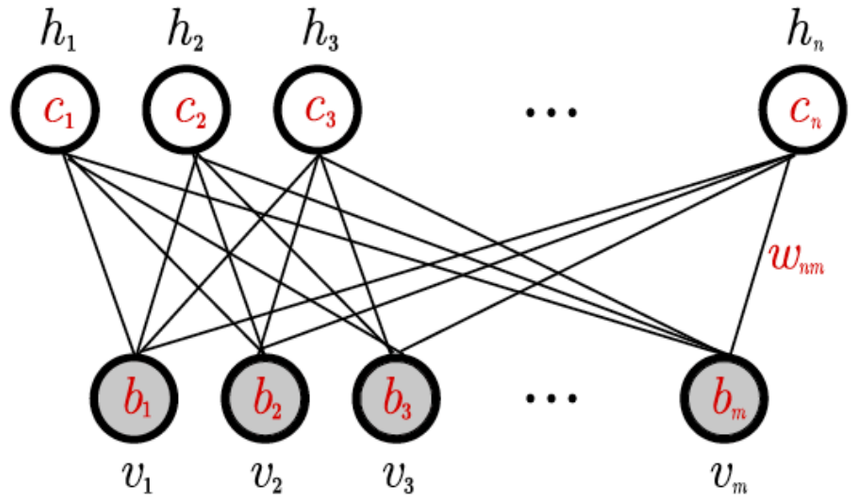
There is no interaction: $I_{i,j}^m = 0$

Spoiler: J_{ij} in Ising!

Recall analytical formula for RBM interactions

$$E_\theta(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{i=1}^{n} c_i h_i - \sum_{j=1}^{m} b_j v_j$$

$$p_{\text{RBM}}(\mathbf{v},\mathbf{h}|\theta) = \frac{1}{Z_{\text{RBM}}} e^{-E_\theta(\mathbf{v},\mathbf{h})}$$



Marginal: $$p(\mathbf{v}|\theta) = \frac{1}{\mathcal{Z}(\theta)} \prod_{j=1}^{m} \left(e^{b_j v_j}\right) \prod_{i=1}^{n} \left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} v_j}\right)$$

Asymptotic expansion, resummation, ...

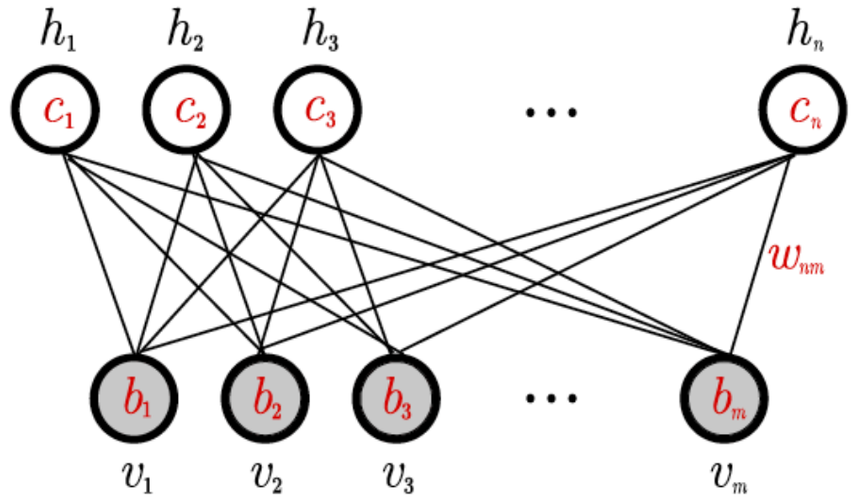
Analytical closed-form expression for n-point interactions, e.g. 2-point:

$$J_{j_1,j_2} \propto \ln \prod_{i=1}^{n} \frac{(1 + e^{c_i + w_{ij_1} + w_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + w_{ij_1}})(1 + e^{c_i + w_{ij_2}})}$$

Recall analytical formula for RBM interactions

$$E_\theta(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{i=1}^{n} c_i h_i - \sum_{j=1}^{m} b_j v_j$$

$$p_{\text{RBM}}(\mathbf{v},\mathbf{h}|\theta) = \frac{1}{Z_{\text{RBM}}} e^{-E_\theta(\mathbf{v},\mathbf{h})}$$



Marginal: $$p(\mathbf{v}|\theta) = \frac{1}{\mathcal{Z}(\theta)} \prod_{j=1}^{m} \left(e^{b_j v_j}\right) \prod_{i=1}^{n} \left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} v_j}\right)$$

Instead, use the TL formulation to directly read-off the coupling!!

$$I^m_{j_1,j_2} = \frac{p(v_{j_1 j_2} = (1,1), \underline{v} = 0)}{p(v_{j_1 j_2} = (1,0), \underline{v} = 0)} \frac{p(v_{j_1 j_2} = (0,0), \underline{v} = 0)}{p(v_{j_1 j_2} = (0,1), \underline{v} = 0)} = \prod_{i=1}^{n} \frac{(1 + e^{c_i + w_{ij_1} + w_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + w_{ij_1}})(1 + e^{c_i + w_{ij_2}})}$$

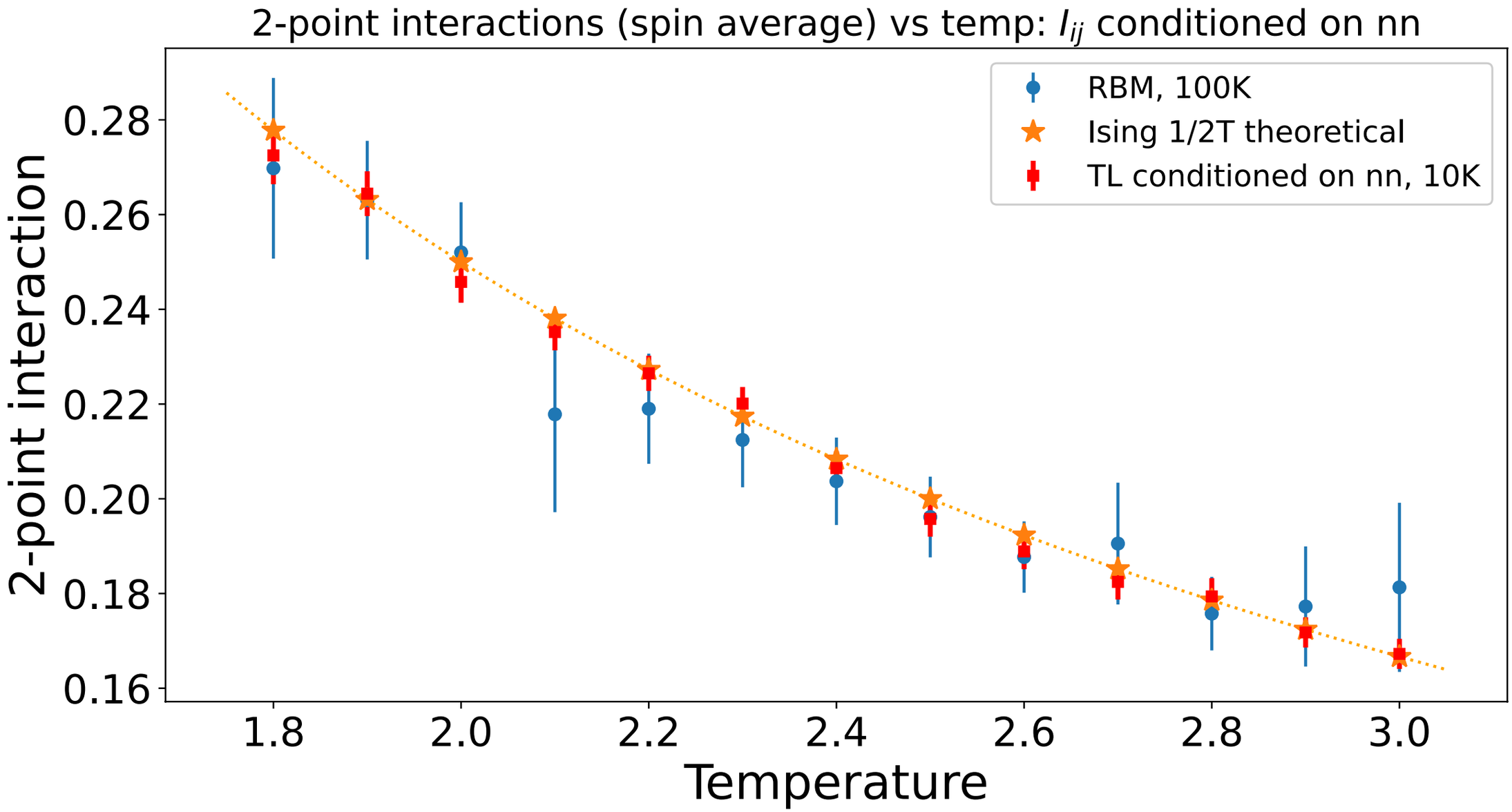
No asymptotic expansion and re-summation required ...

Applies to other energy based models

Beentjes & Khamseh, Physical Review E (2020)

Model-independent estimation results

Conditioning on parent spins to isolate pairs from the rest of the system (Markovian). Run time: Few seconds per temperature.



10K samples

Biology: Large number of dependent variables

Estimating intricate interaction structure amongst many genes

Certain approximation no longer possible: $p(G_i, G_j) \neq p(G_i)p(G_j)$

Number of variables >> data, (and high temperatures) **G binarised!**

$$I^m_{i,j} = \ln\left(\frac{p(G_{ij} = (1,1) \mid \underline{G} = 0)}{p(G_{ij} = (0,1) \mid \underline{G} = 0)} \frac{p(G_{ij} = (0,0) \mid \underline{G} = 0)}{p(G_{ij} = (1,0) \mid \underline{G} = 0)}\right)$$

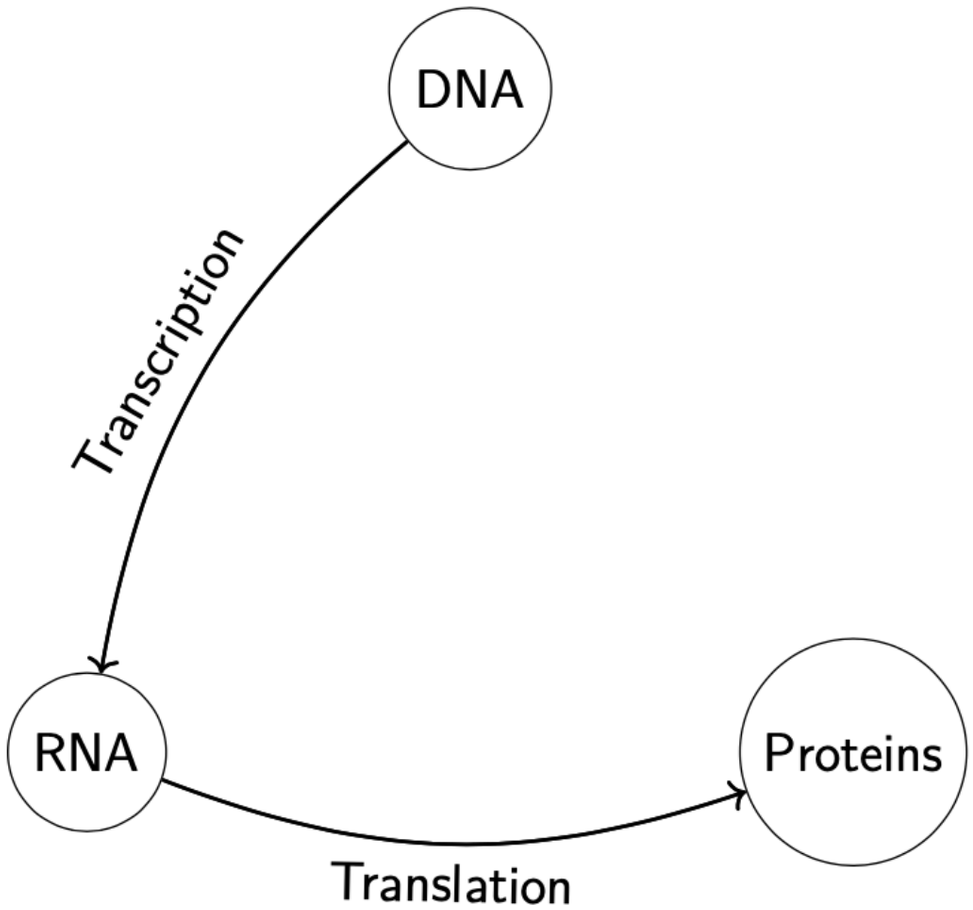
Estimate conditional dependencies directly from data, using efficient causal discovery algorithms (e.g. PC, Score-based MCMC)

Nothing comes for free! These come with their own assumptions/bias
Keep in mind to be conservative.

Part 3

Biological data: Gene expression

The central dogma



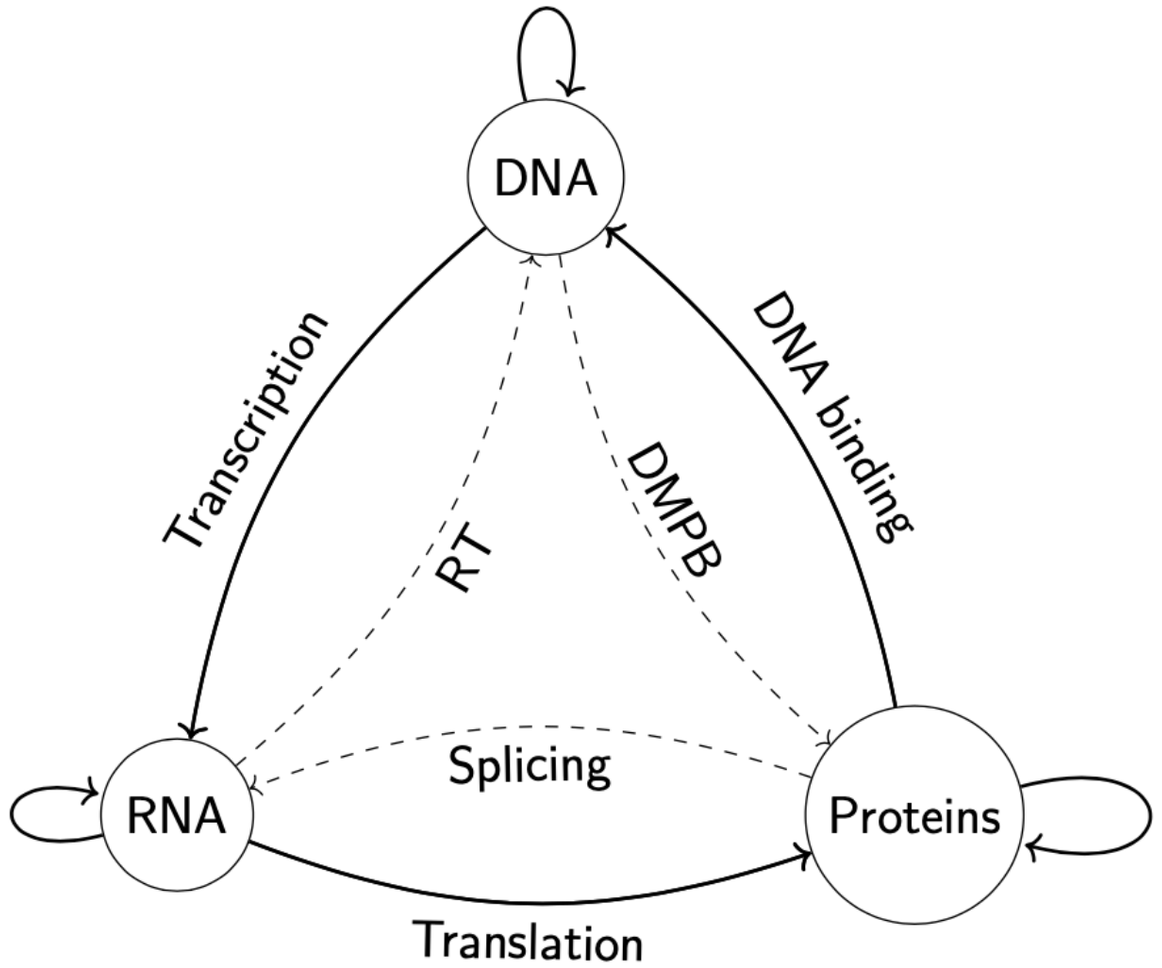
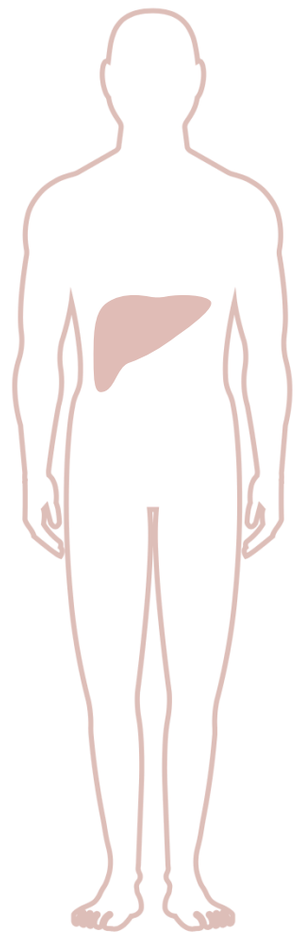
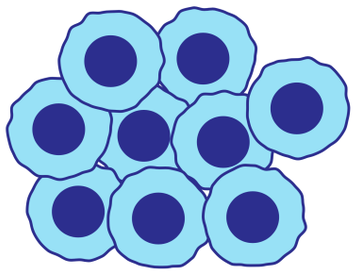


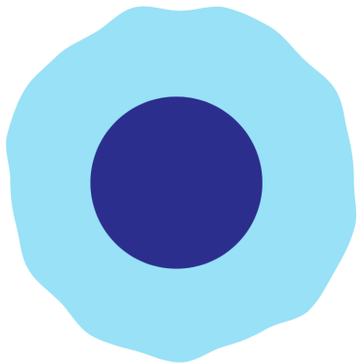
Figure by Abel Jansma (PhD student)

Biology: Large number of dependent variables

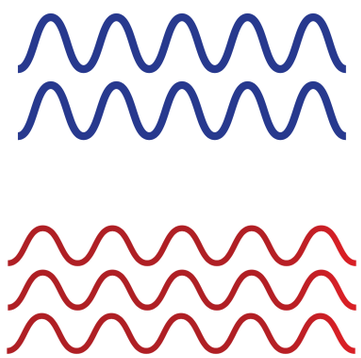




Single Cell



RNA



Sequencing

Illumina

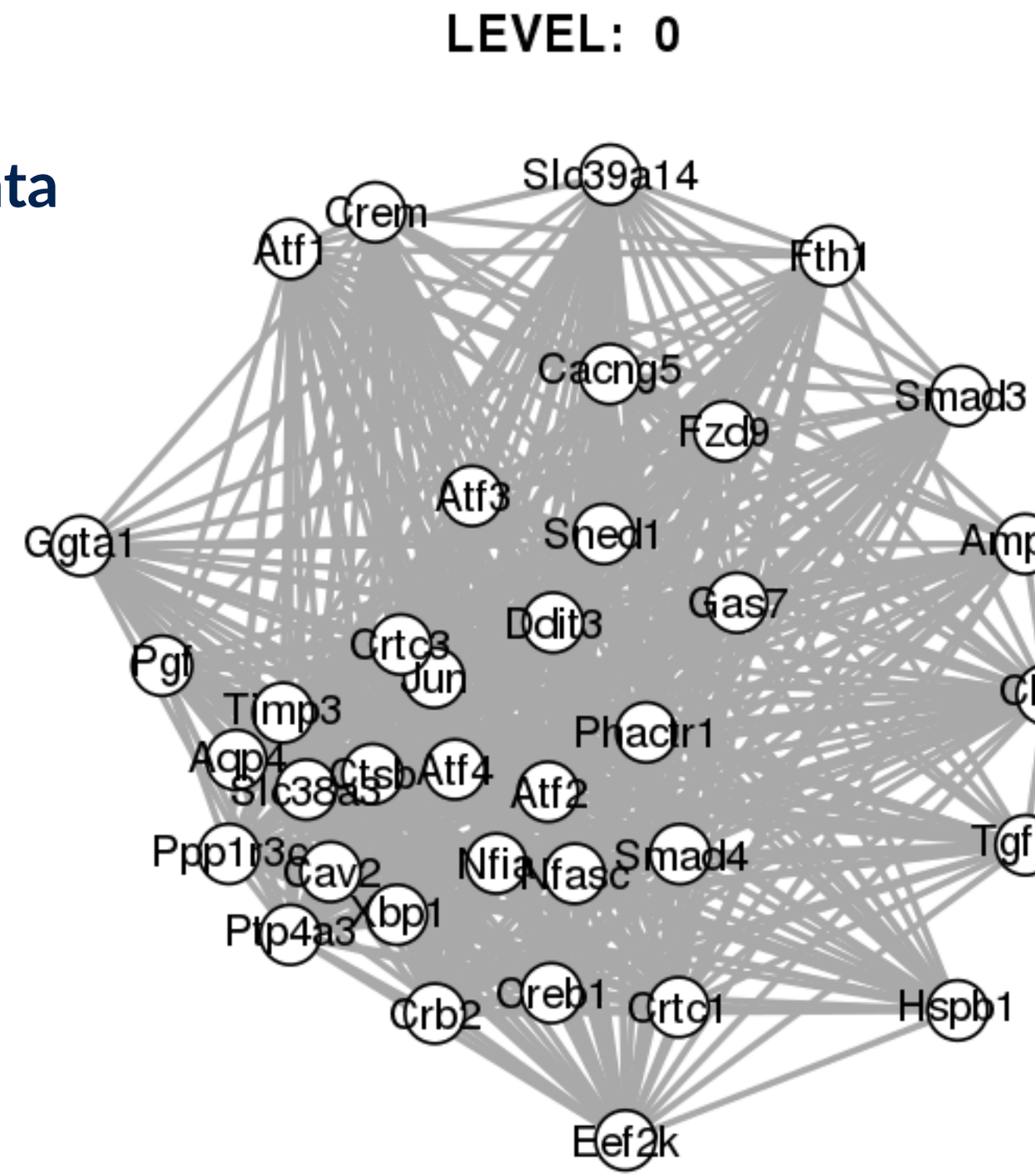
Count data
(Later binarised)

Gene

Cell [ ] NxG

(In)dependence of gene expressions

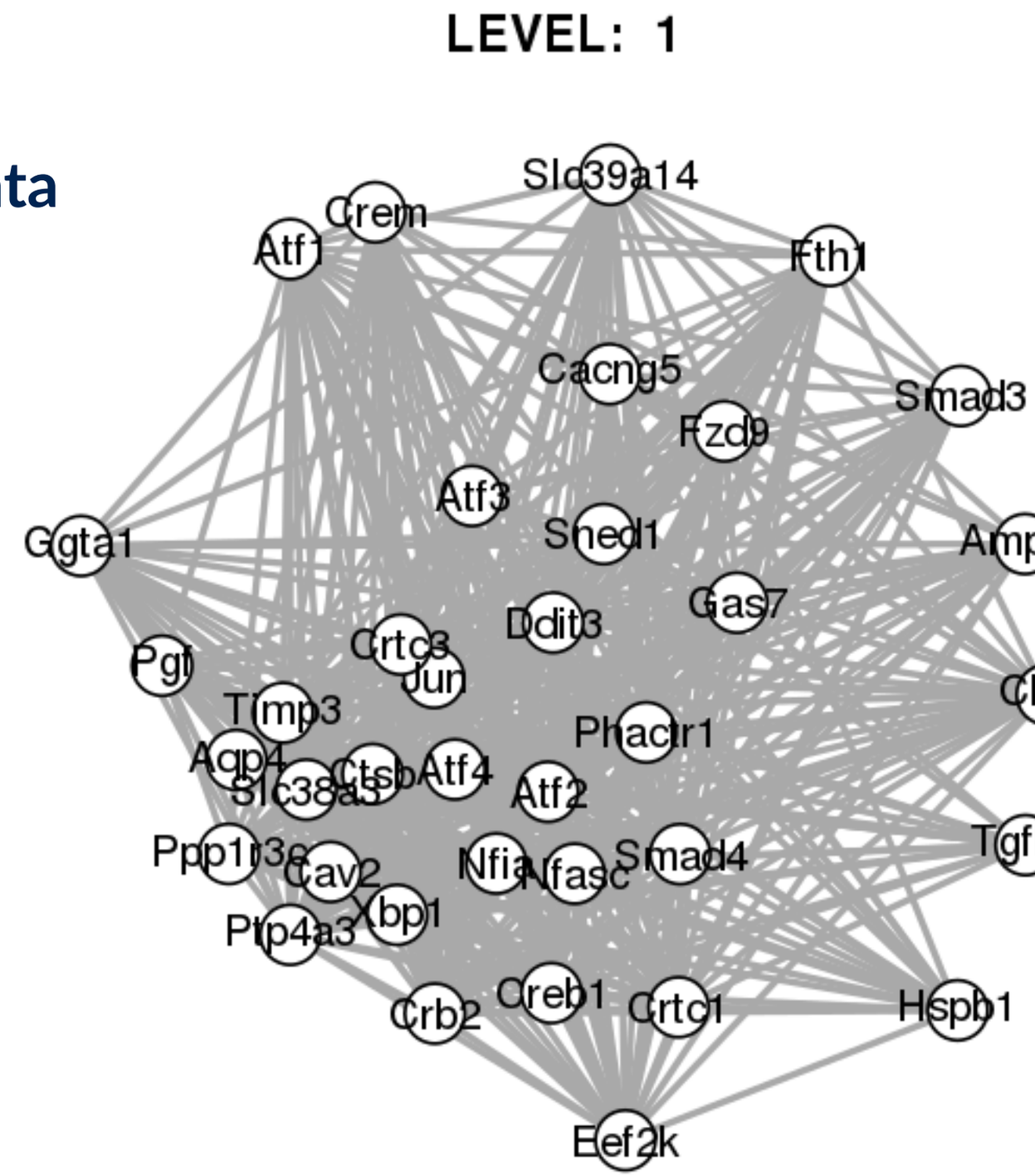
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

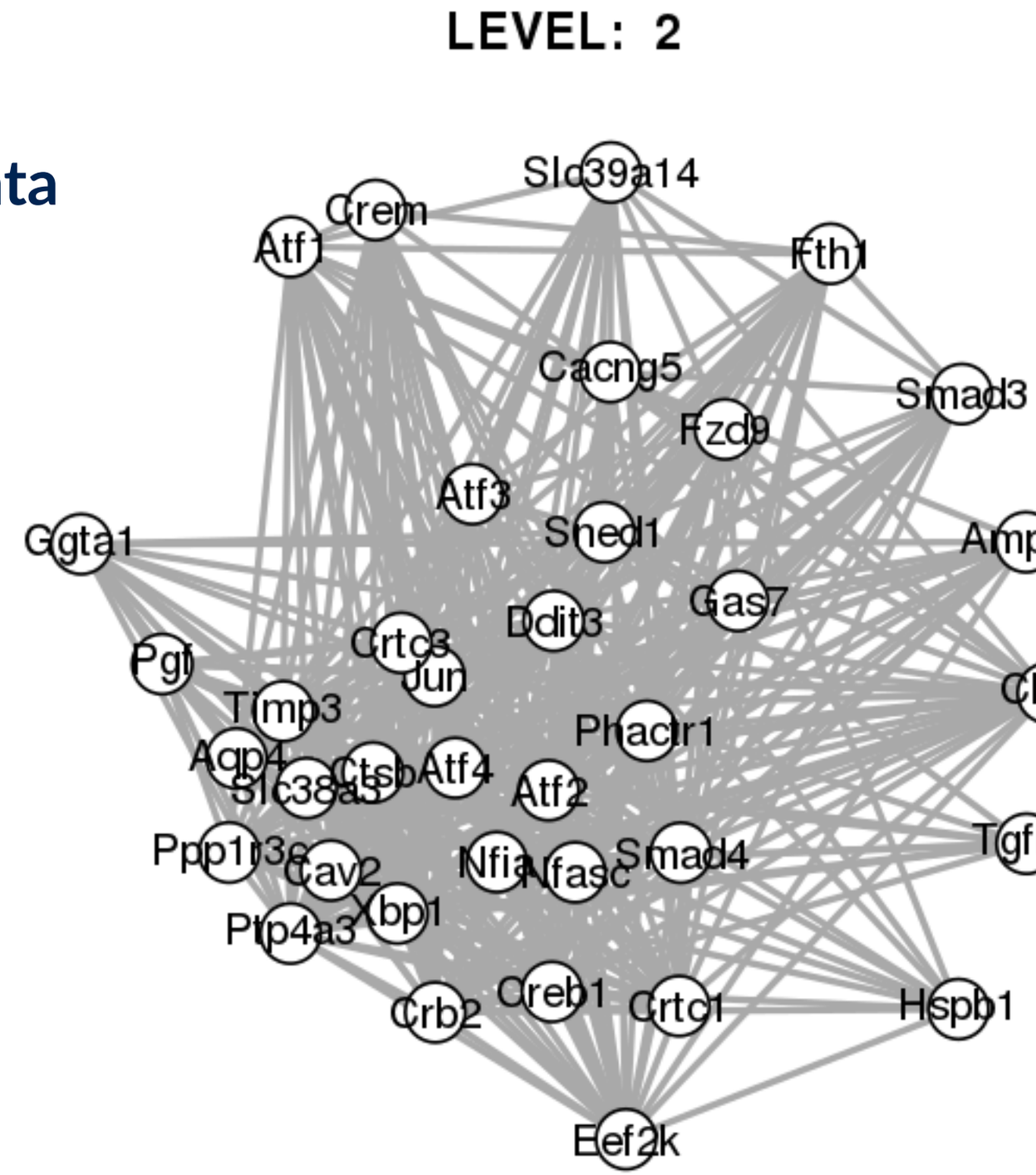
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

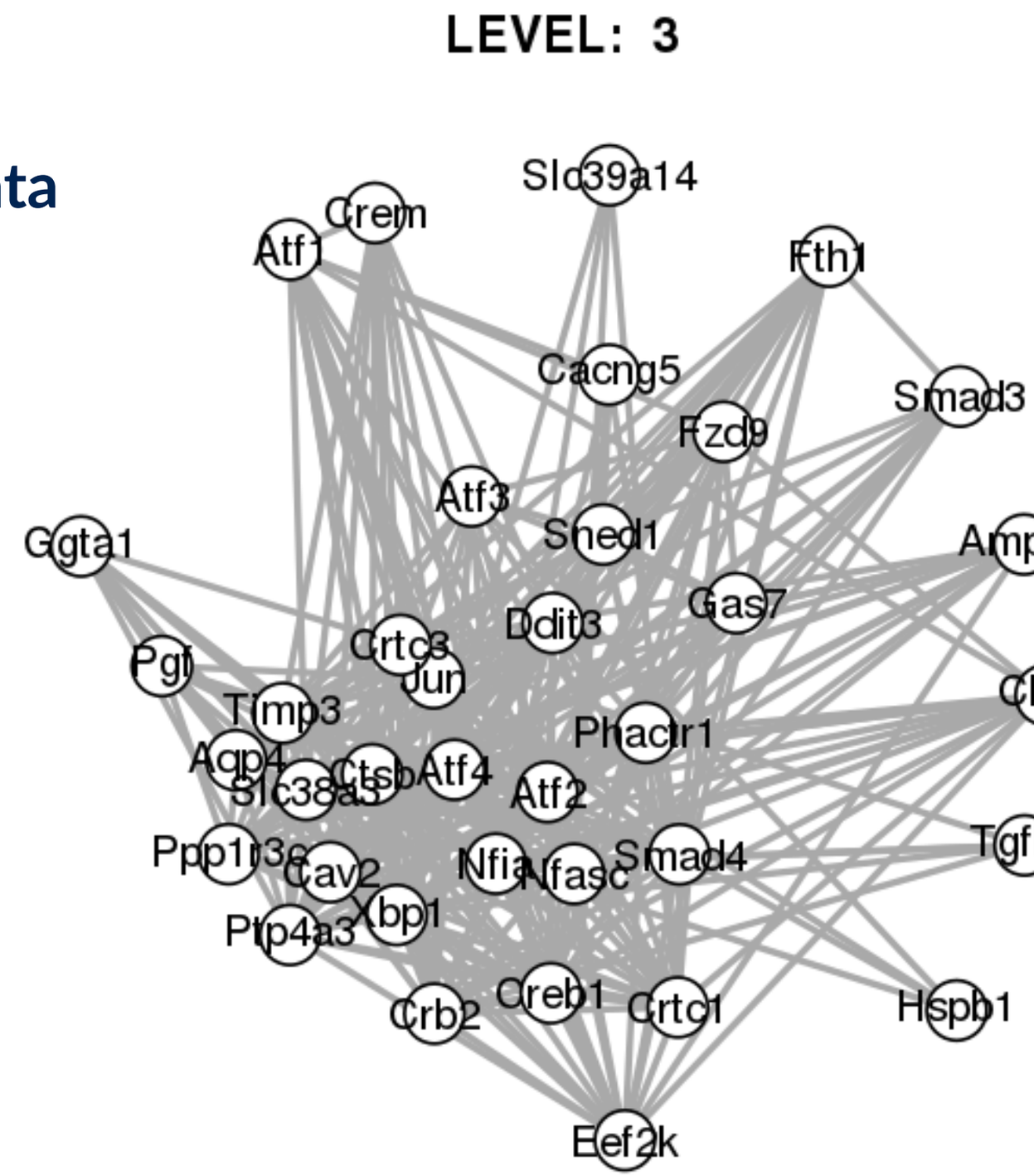
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

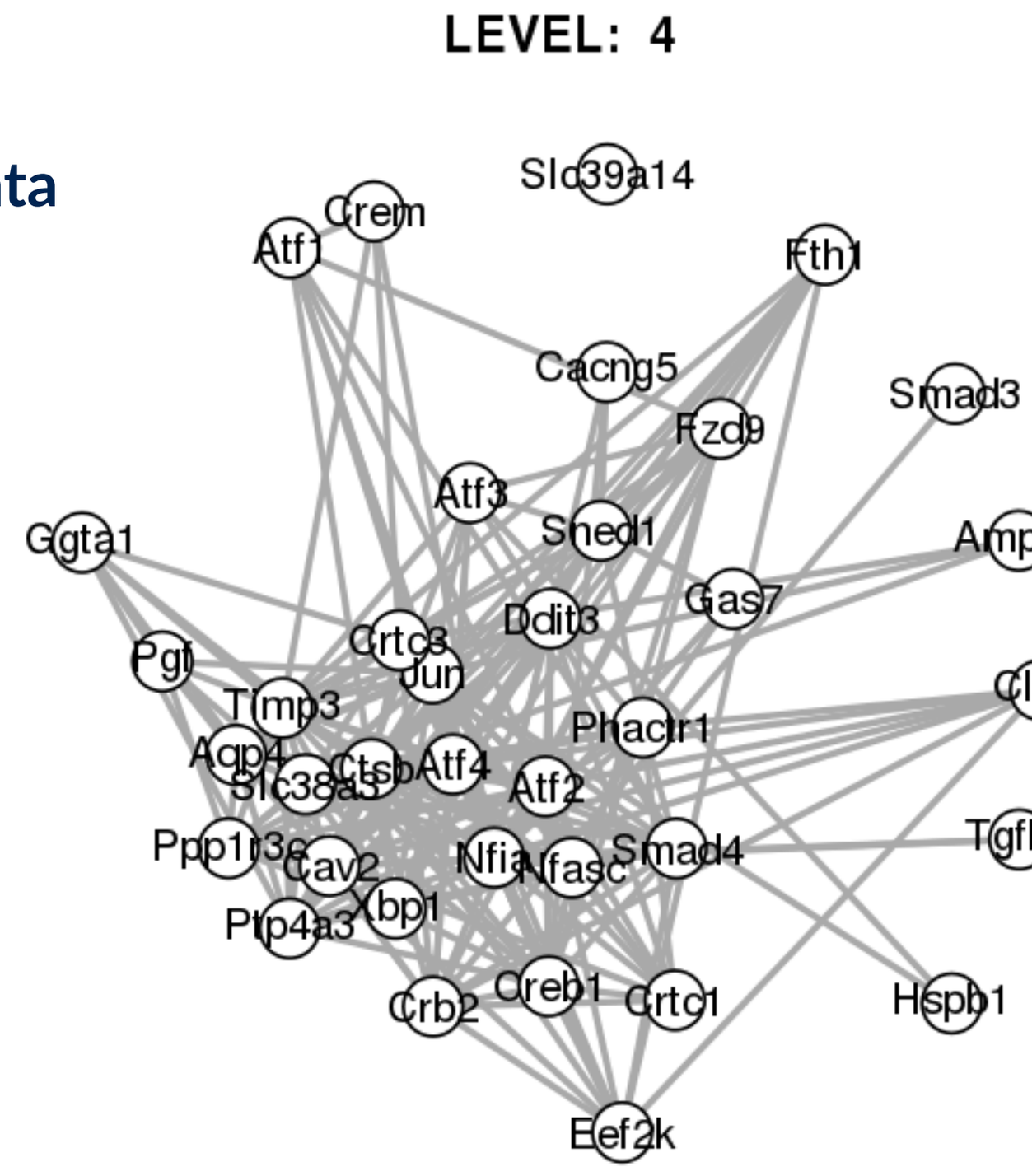
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

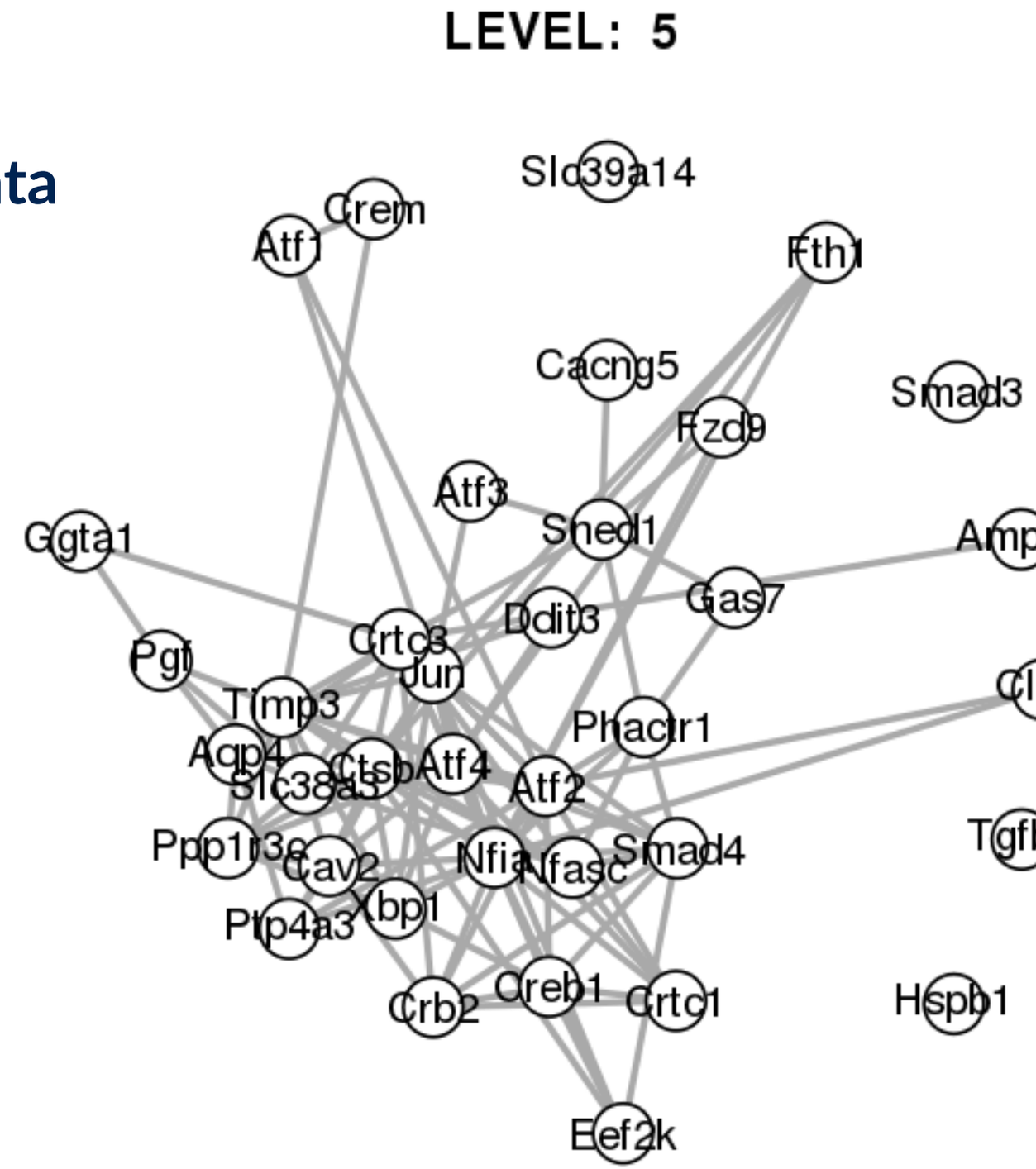
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

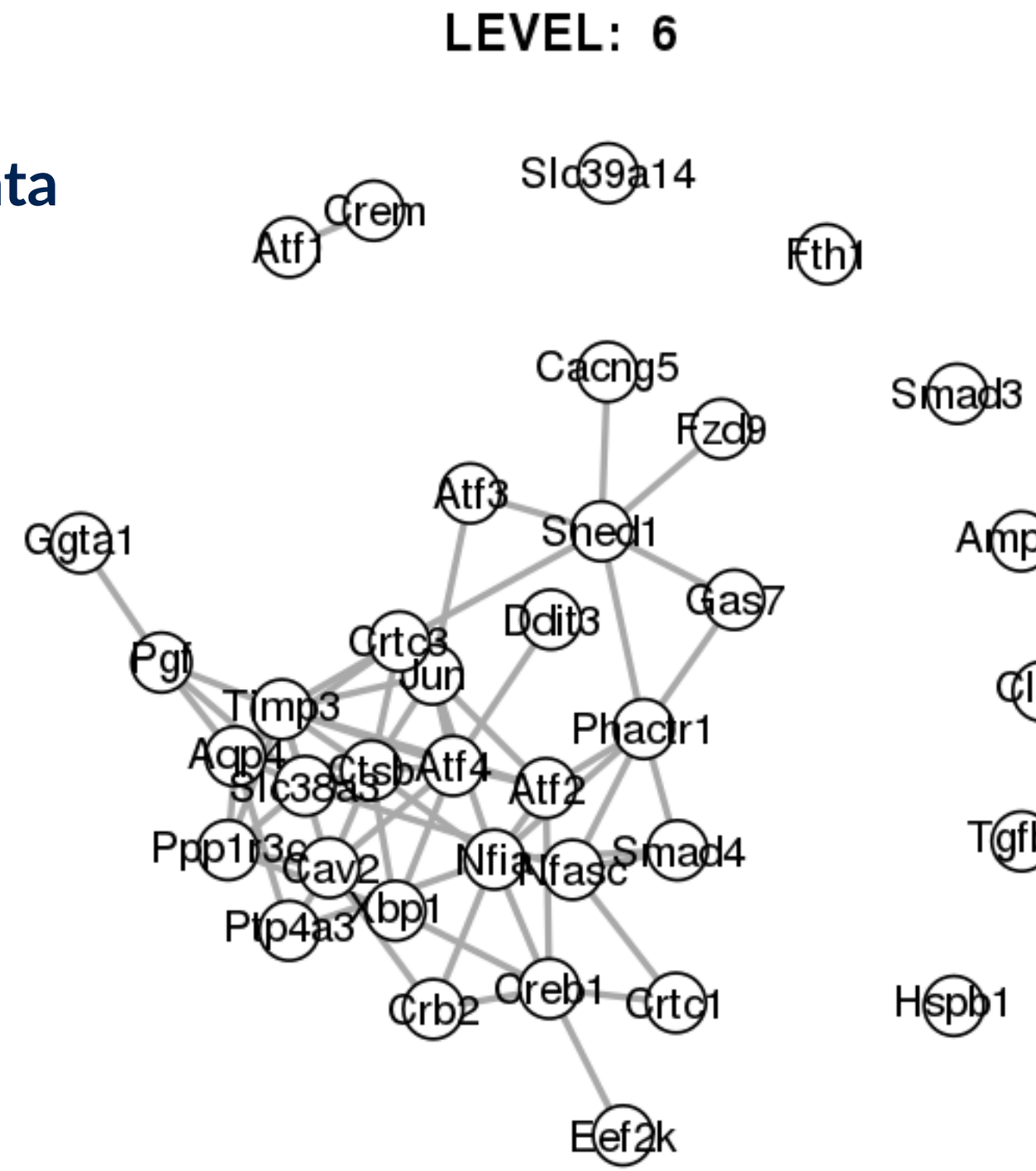
10X single-cell
1M mouse brain
developmental data



Plot by Abel Jansma (PhD student)

(In)dependence of gene expressions

10X single-cell
1M mouse brain
developmental data

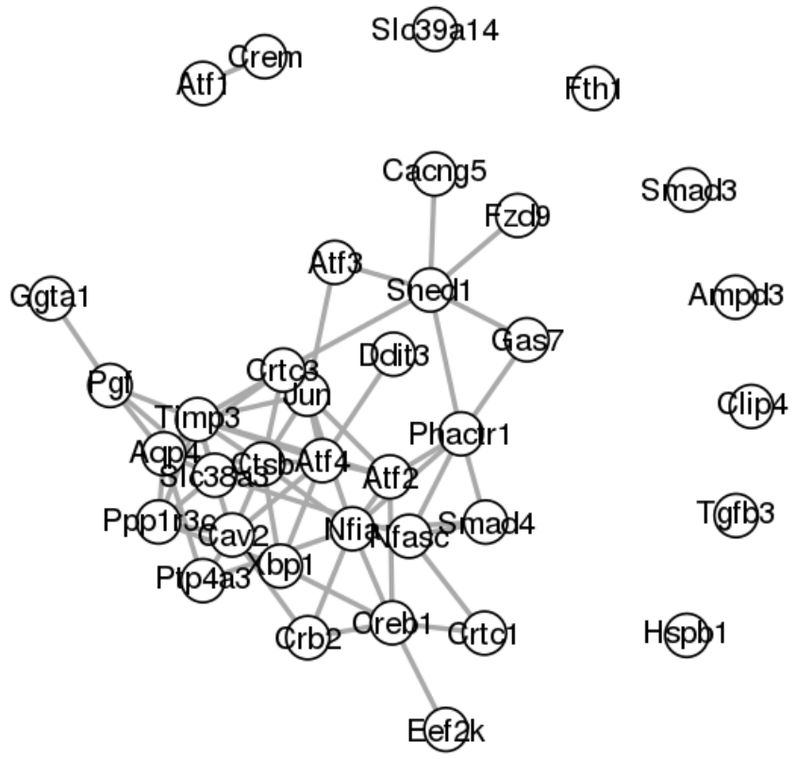


Plot by Abel Jansma (PhD student)

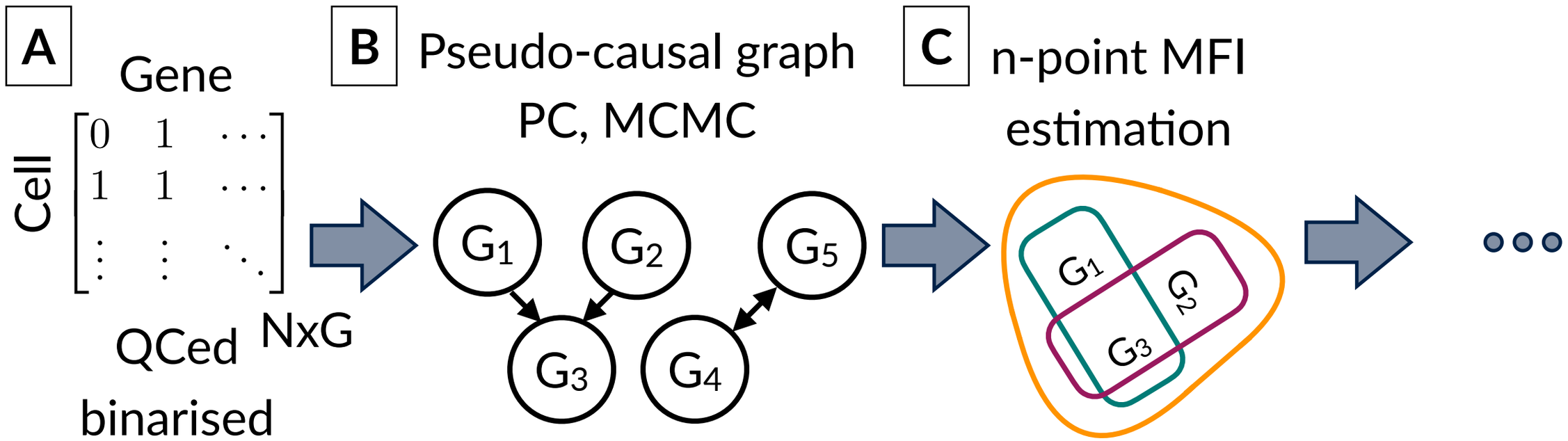
Estimate model-free n-point interactions

$$I^m_{i,j} = \ln\left(\frac{p(G_{ij}=(1,1)\mid \underline{G}=0)}{p(G_{ij}=(0,1)\mid \underline{G}=0)}\frac{p(G_{ij}=(0,0)\mid \underline{G}=0)}{p(G_{ij}=(1,0)\mid \underline{G}=0)}\right)$$

2-point up to 7-point interactions



Stator workflow



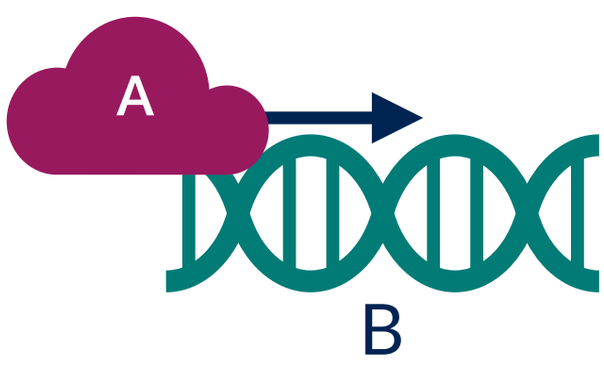
What is the biological interpretation of these n-point interactions?

Does it help answer questions of molecular & cell biologists?

Regulation vs Cell State

1. MFIs represent physical interactions amongst molecules **within** a cell

2. MFIs represent a biochemical network:
Transcription factor A -> Target gene B



3. MFIs represent dependence structure amongst genes that imply
cell types, subtypes or states
(MFIs estimated as an average across **diverse** cell populations)
i.e. the statistical interpretation of interactions, rather than dynamical/physical

Stator workflow

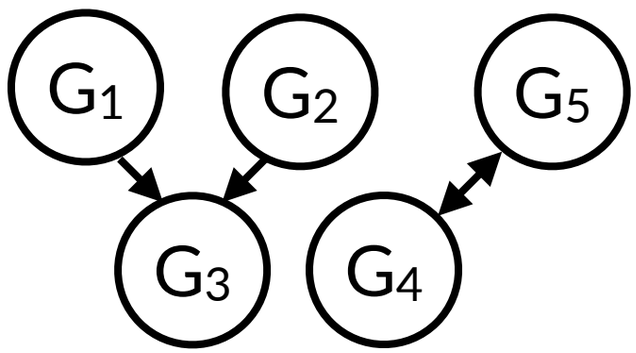
A

Gene

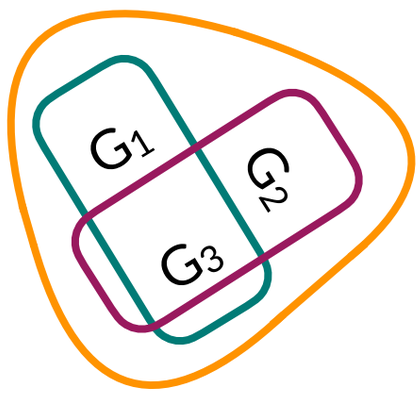
Cell $\begin{bmatrix} 0 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$ NxG

QCed binarised

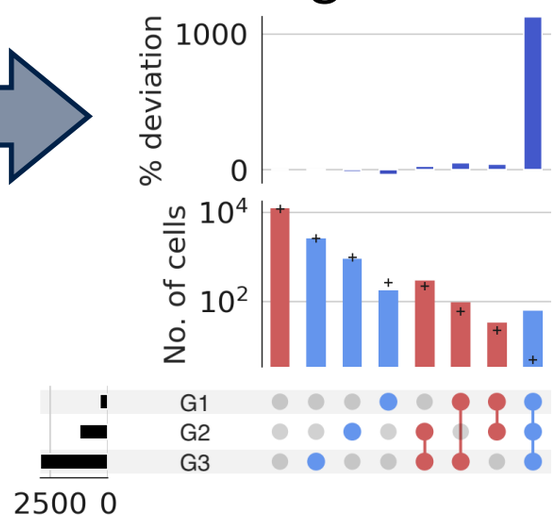
B Pseudo-causal graph PC, MCMC



C n-point MFI estimation



D Deviating tuples (d-tuples) of genes driving MFIs



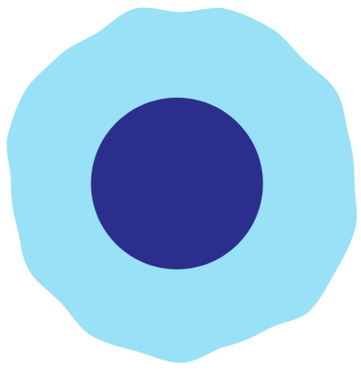
$$\text{d-tuple} : (G_1, G_2, G_3) = (1, 1, 1)$$

E

d-tuple

Cell $\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$ NxK

For each cell



$$\text{if}(G_1, G_2, G_3) = (1, 1, 1) \Rightarrow 1$$
$$\text{if}(G_1, G_2, G_3) \neq (1, 1, 1) \Rightarrow 0$$

Stator pipeline: Abel Jansma (PhD student)

Stator workflow

A

Gene

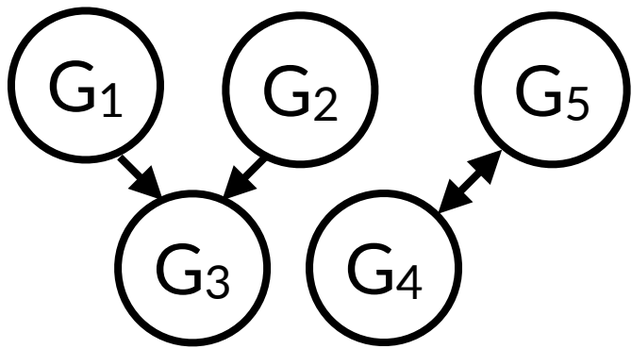
Cell

$$\begin{bmatrix} 0 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

QCed binarised

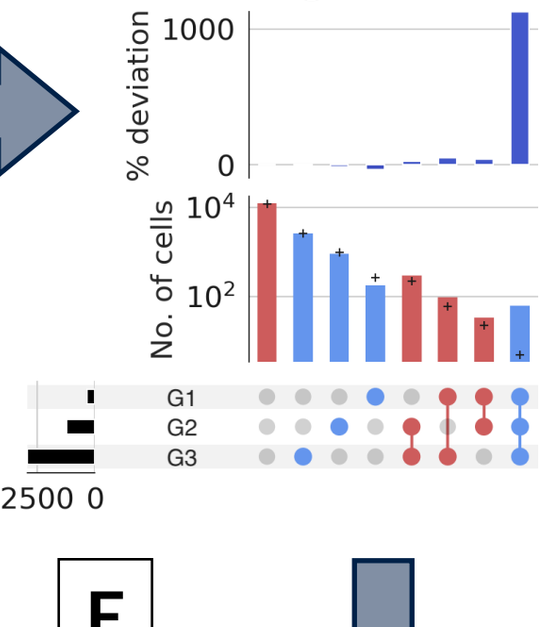
NxG

B Pseudo-causal graph PC, MCMC



C n-point MFI estimation

D Deviating tuples (d-tuples) of genes driving MFIs



E

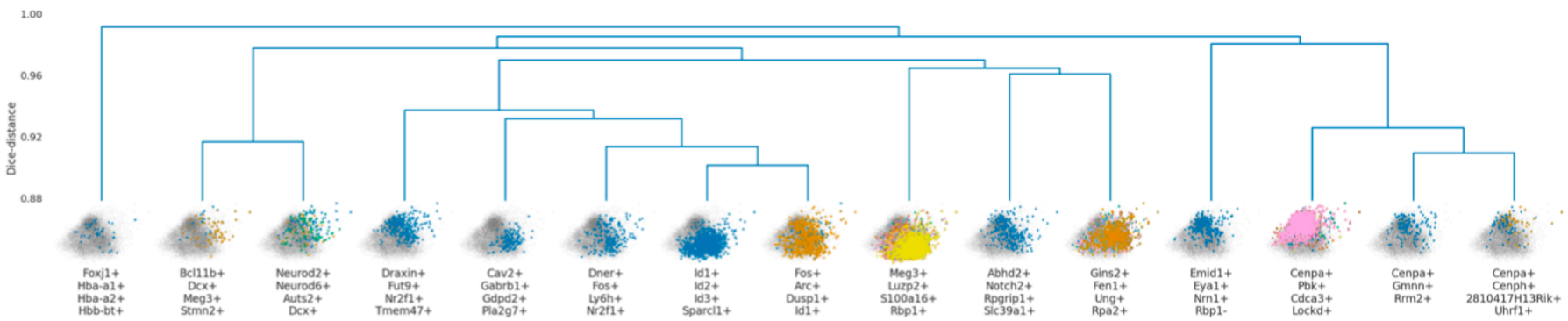
d-tuple

Cell

$$\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

NxK

F Hierarchical Clustering of d-tuples (Visualisation in lower dimensional embedding)



Biological States

Stator pipeline: Abel Jansma (PhD student)

Stator workflow

A

Gene

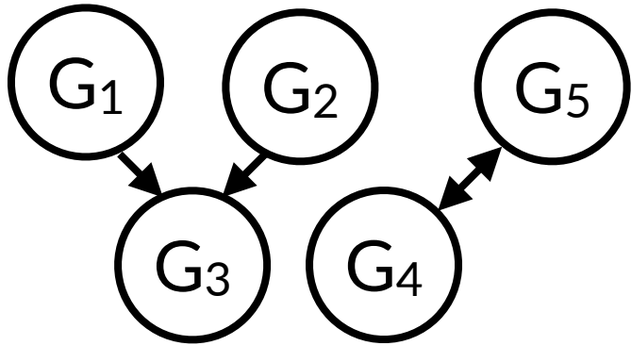
Cell

$$\begin{bmatrix} 0 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

QCed binarised

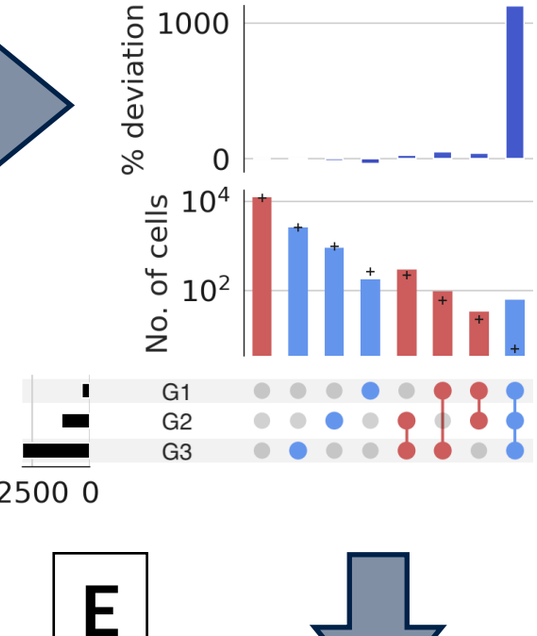
NxG

B Pseudo-causal graph PC, MCMC



C n-point MFI estimation

D Deviating tuples (d-tuples) of genes driving MFIs



E

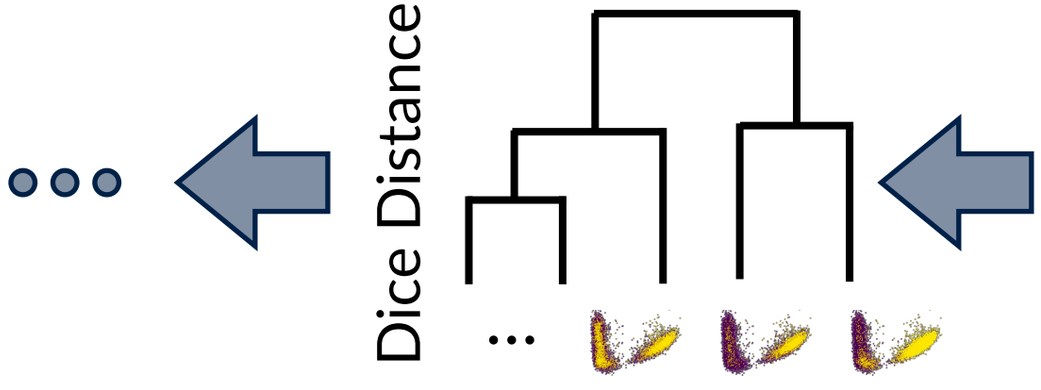
d-tuple

Cell

$$\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

NxK

F Hierarchical Clustering of d-tuples



Biological interpretation of **states** by combining d-tuples at a given dice distance

Will biologists bother using this methodology?

Criteria:

1. Ease of use
   (easy-to-follow documentation, easy copy/paste code, press of a button)
2. Speed
3. Output with good visualisation
4. Biology that “makes sense”

...

...

...

Criteria:

1. Ease of use
   (easy-to-follow documentation, easy copy/paste code, press of a button)
2. Speed
3. Output with good visualisation
4. Biology that “makes sense”

…

…

…

n. The methodology behind the software

(where n can be very large!)

Stator workflow

A

Gene

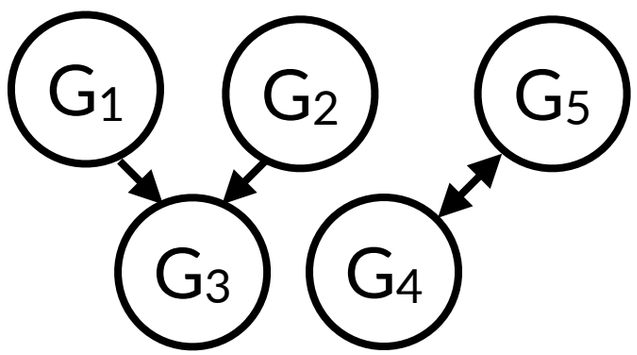
Cell

$$\begin{bmatrix} 0 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

QCed binarised

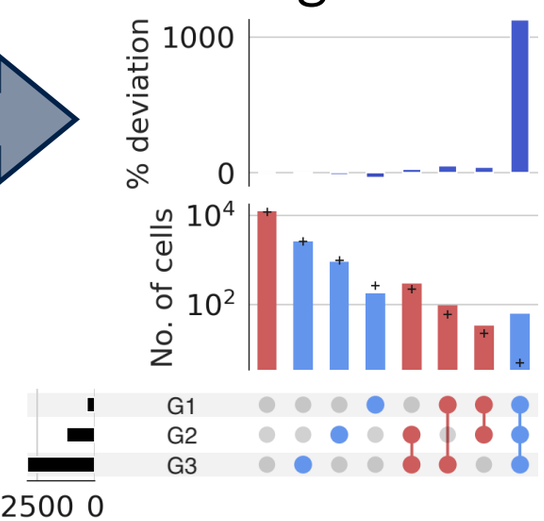
NxG

B Pseudo-causal graph PC, MCMC

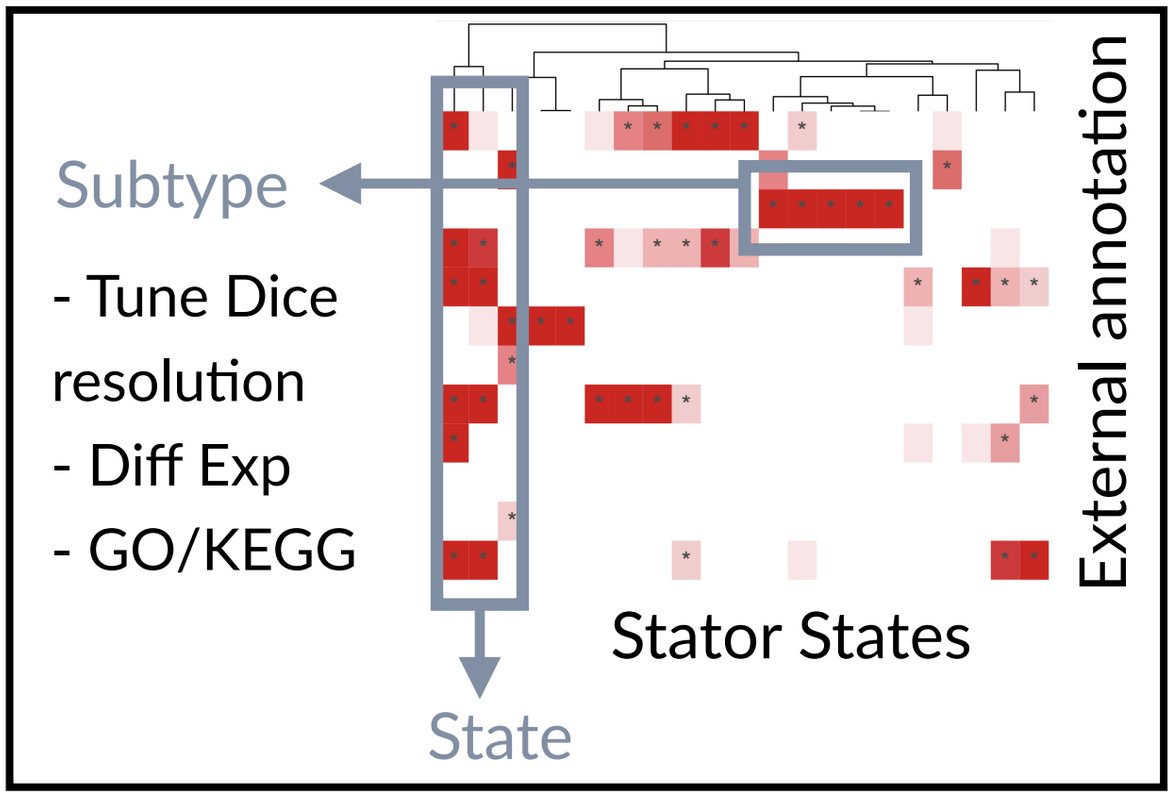


C n-point MFI estimation

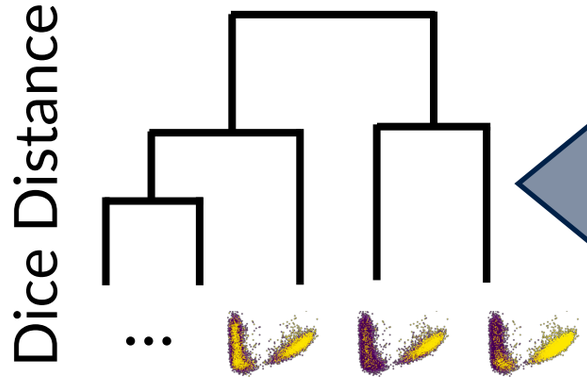
D Deviating tuples (d-tuples) of genes driving MFIs



G Downstream Analysis with Interactive Shiny App in R



F Hierarchical Clustering of d-tuples



E

d-tuple

Cell

$$\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

NxK

Biological interpretation of **states** by combining d-tuples at a given dice distance

Liver Cancer: Cell types and states

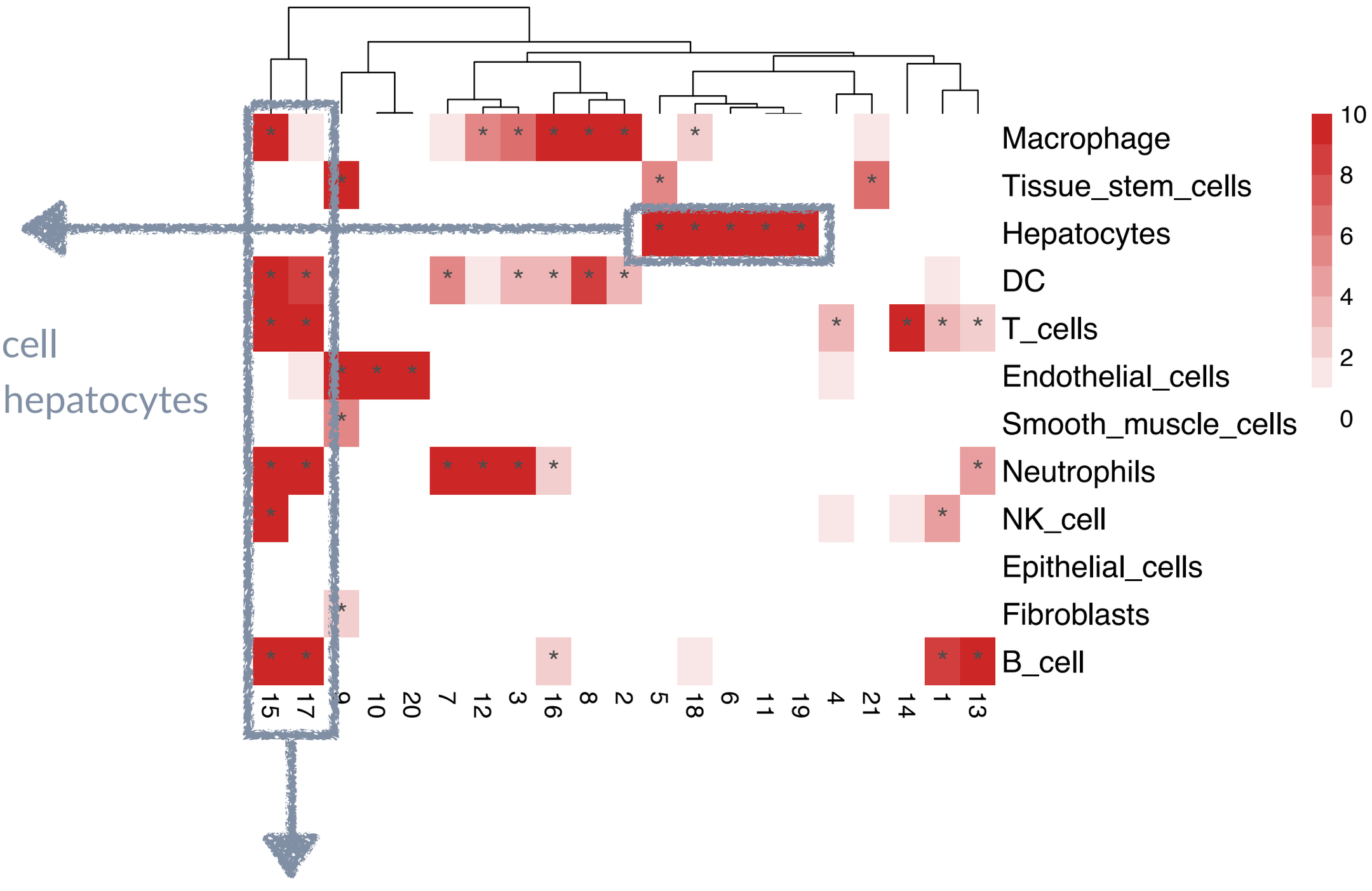
Subtype:

Fibrotic

cancer stem cell

proliferative hepatocytes

...



State: e.g. cell cycle, proliferative states, high-energy state, ...

Stator app interface

If you like MFIs and use it, please consider citing the related article:
Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium Sjoerd Viktor Beentjes and Ava Khamseh, Phys. Rev. E. 2020 Nov 102, 053314

MFIs application code is available through Github

If you have any question, you can send an e-mail

About | Table | Heatmap | GO & KEGG | Using rrvgo | Upset Plot | DE analysis

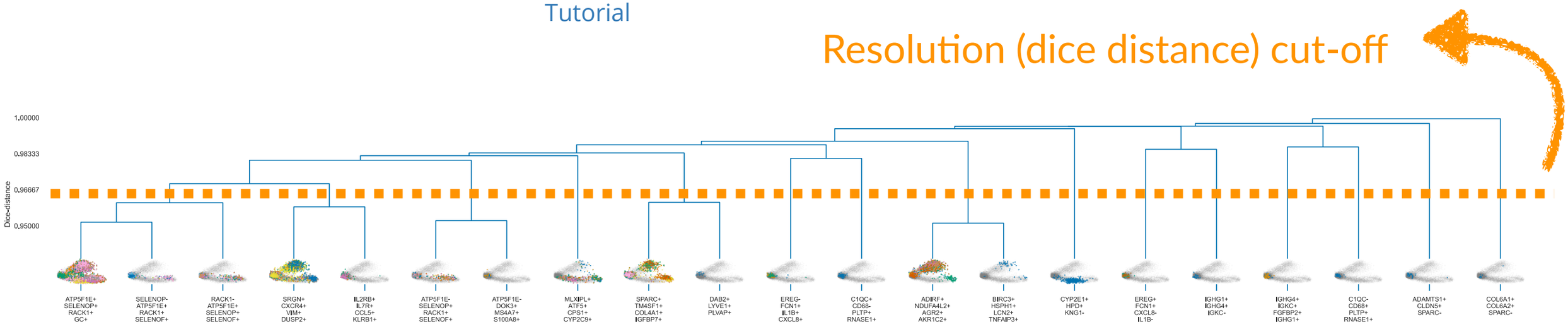
Explore cell states by MFIs

MFIs takes in scRNA-seq count matrix and estimate gene interactions. Here we show how to use these MFIs to explore cell states.

Data Visualization & Analysis

- Table - A Summary statistics for deviating state
- Heatmaps - Over-representation test for MFIs and other cell annotations
- GO & KEGG for genes in each state
- rrvgo - Simplifying the redundance of GO sets
- Upset Plot
- DE analysis for mutually exclusive states

Tutorial



Stator interface app: Yuelin Yao (PhD student)

Stator app interface

Here we use scRNA-seq HCC dataset. To upload your data, click the box:

ON

Dice distance:

0 0.95

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Submit

Resolution cut-off

About | Table | Heatmap | GO & KEGG | Using rrvgo | Upset Plot | DE analysis

Table for deviating MFIs

391 deviating MFIs in total, 21 clusters.

Show 10 entries

Search:

	genes	state	dev	pval	cluster
47	IGHG4_IGHG1_IGKC	111	216.09886998247	2.29736631083133e-28	Cluster:1
150	IGHG4_FGFBP2_IGKC	111	21.9122324209873	1.04827895292059e-19	Cluster:1
341	PLTP_CD68_C1QC_RNASE1	1101	6.41225201584546	1.39266594426104e-9	Cluster:2
213	EREG_CXCL8_IL1B_FCN1	1011	14.7648630077289	1.05933261179815e-8	Cluster:3
220	EREG_CXCL8_IL1B_FCN1	1101	13.9465435194403	3.12038785771356e-13	Cluster:3
282	EREG_IL1B_FCN1	101	9.95679034074909	2.46844896187663e-13	Cluster:3
292	EREG_CXCL8_FCN1	101	9.09789155019342	9.41630198774148e-13	Cluster:3
177	IGHG4_IGHG1_IGKC	110	18.5991179035581	8.521545688936e-16	Cluster:4
63	CXCL1_IER3_CXCL3	111	99.0782359764262	2.01516836396423e-47	Cluster:5
74	MB_NDUFA4L2_TFF2	111	79.5037037037037	6.82448451177721e-18	Cluster:5

Showing 1 to 10 of 391 entries

Previous 1 2 3 4 5 ... 40 Next

Download as .csv

Stator interface app: Yuelin Yao (PhD student)

Stator app interface

Compare to known cell types from the literature

Compare to other methods used in the literature, e.g. NMF

Compare to expert annotated cell types

Stator interface app: Yuelin Yao (PhD student)

Stator workflow

A

Gene

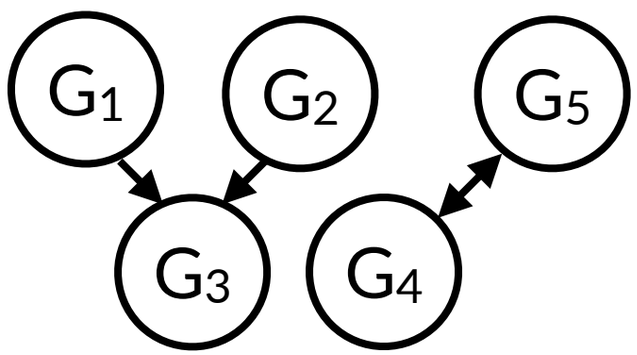
Cell

$$\begin{bmatrix} 0 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

QCed binarised

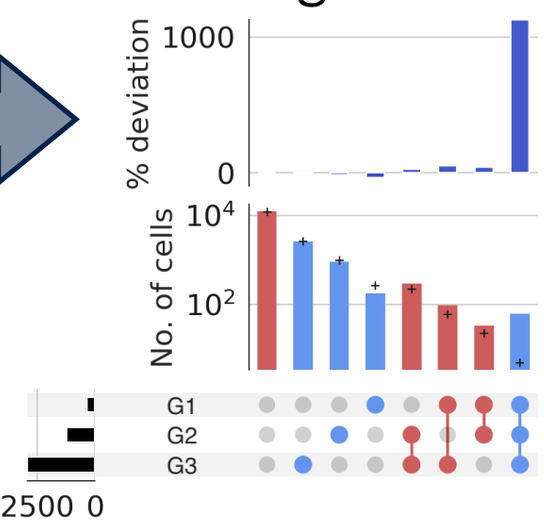
NxG

B Pseudo-causal graph PC, MCMC

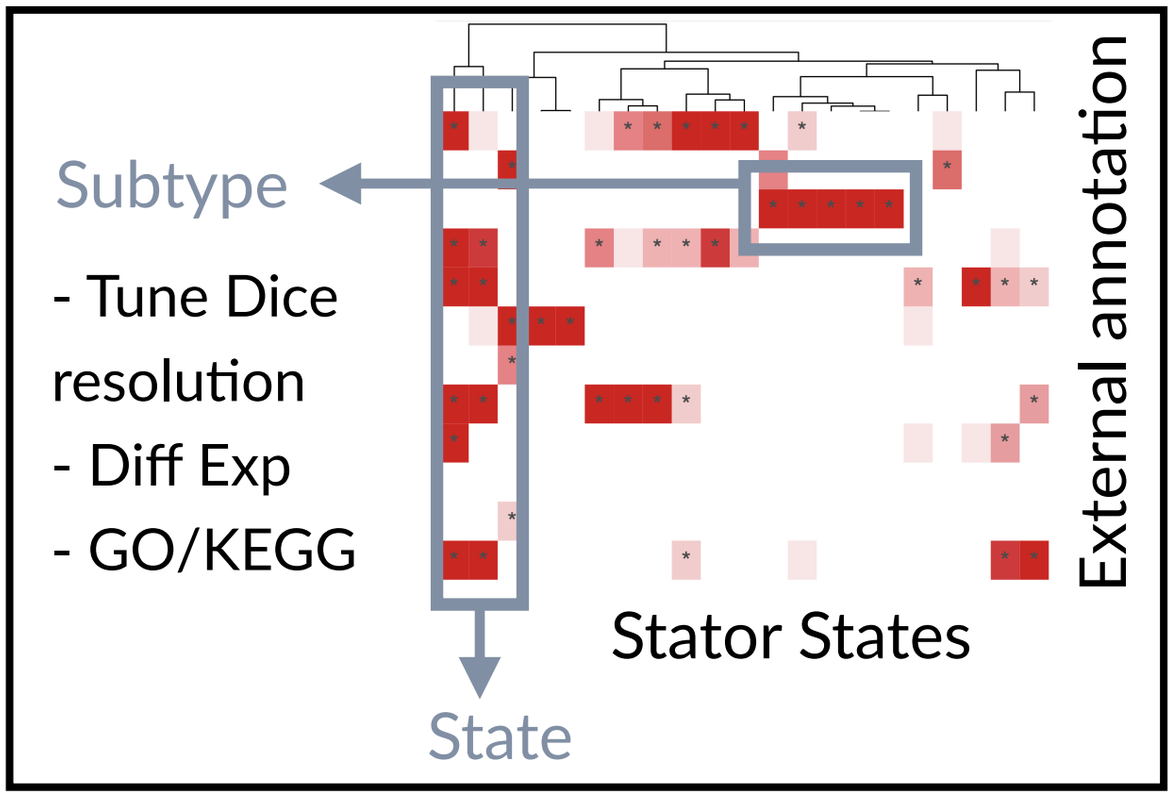


C n-point MFI estimation

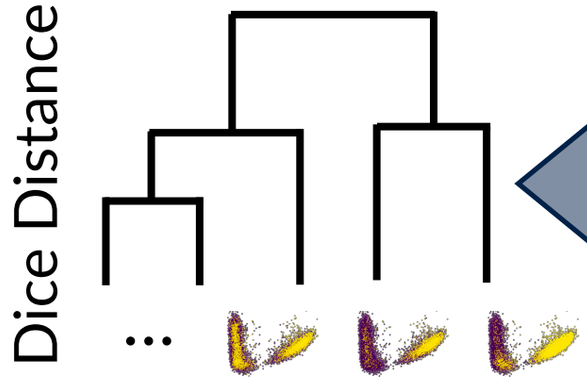
D Deviating tuples (d-tuples) of genes driving MFIs



G Downstream Analysis with Interactive Shiny App in R



F Hierarchical Clustering of d-tuples



E

d-tuple

Cell

$$\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

NxK

Biological interpretation of **states** by combining d-tuples at a given dice distance

Backup slides

RBM Prediction: n-point interactions

- A non pair-wise treatment
- Higher order couplings
- Not accessible via standard statistical techniques

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_j \left(\sum_i \kappa_i^{(i)} W_{ij} \right) v_j - \frac{1}{2} \sum_{jk} \underbrace{\left(\sum_i \kappa_i^{(2)} W_{ik} W_{ij} \right)}_{} v_j v_k + \cdots$$

Re-sum the entire series to obtain 2-point coupling!!

Derivation of n-point interactions in closed form

$$E(\mathbf{v}) = \ln \sum_{\mathbf{h}} e^{E(\mathbf{v},\mathbf{h})}$$

$$= \ln \sum_{\mathbf{h}} e^{-\sum_j b_j v_j - \sum_i c_i h_i - \sum_{i,j} h_i W_{ij} v_j}$$

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_i \ln \sum_{h_i} e^{c_i h_i} e^{\sum_j h_i W_{ij} v_j}$$

$$= -\sum_j b_j v_j - \sum_i \ln \sum_{h_i} q(h_i) e^{t h_i} \ , \qquad t \equiv \sum_j W_{ij} v_j \text{ and } q(h_i) \equiv e^{c_i h_i}$$

Cumulant generating function:

$$K_i(t) \equiv \ln \sum_{h_i} q(h_i) e^{t h_i} = \sum_n \frac{\kappa_i^{(n)} t^n}{n!} \qquad \kappa_i^{(n)} = \partial_t^n K_i(t)|_{t=0}$$

A high-bias, low-variance introduction to machine learning for physicists by Mehta et al. *Physics Reports* (2019)

Derivation of n-point interactions in closed form

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_i \kappa_i^{(0)} - \sum_i \kappa_i^{(1)} t - \sum_i \frac{\kappa_i^{(2)} t^2}{2!} - \dots$$

$$= -\sum_i \kappa_i^{(0)} - \sum_j \left(b_j + \sum_i \kappa_i^{(1)} W_{ij} \right) v_j - \frac{1}{2!} \sum_{j_1, j_2} \left(\sum_i \kappa_i^{(2)} W_{ij_1} W_{ij_2} \right) v_{j_1} v_{j_2} - \dots$$

$$\boxed{v_j^n = v_j \quad , \quad n \in \mathbb{Z}^+}$$

e.g. 2-point interaction:

$$\sum_{n>1} \frac{1}{2(n!)} \sum_{0<k<n} \sum_{j_1 \neq j_2} \left(\sum_i \kappa_i^{(n)} \binom{n}{k} W_{ij_1}^k W_{ij_2}^{n-k} \right) v_{j_1} v_{j_2}$$

...

$$H_{j_1 j_2} = \frac{1}{8} \sum_i \ln \frac{(1 + e^{c_i + W_{ij_1} + W_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + W_{ij_1}})(1 + e^{c_i + W_{ij_2}})}$$

Closed form expression!

Cossu et. al., Physical Review B (2018)

Derivation of n-point interactions in closed form

$$E(\mathbf{v}) = -\sum_j b_j v_j - \sum_i \kappa_i^{(0)} - \sum_i \kappa_i^{(1)} t - \sum_i \frac{\kappa_i^{(2)} t^2}{2!} - \dots$$

$$= -\sum_i \kappa_i^{(0)} - \sum_j \left(b_j + \sum_i \kappa_i^{(1)} W_{ij} \right) v_j - \frac{1}{2!} \sum_{j_1, j_2} \left(\sum_i \kappa_i^{(2)} W_{ij_1} W_{ij_2} \right) v_{j_1} v_{j_2} - \dots$$

$$\boxed{v_j^n = v_j \quad , \quad n \in \mathbb{Z}^+}$$

e.g. 2-point interaction:

$$\sum_{n>1} \frac{1}{2(n!)} \sum_{0<k<n} \sum_{j_1 \neq j_2} \left(\sum_i \kappa_i^{(n)} \binom{n}{k} W_{ij_1}^k W_{ij_2}^{n-k} \right) v_{j_1} v_{j_2}$$

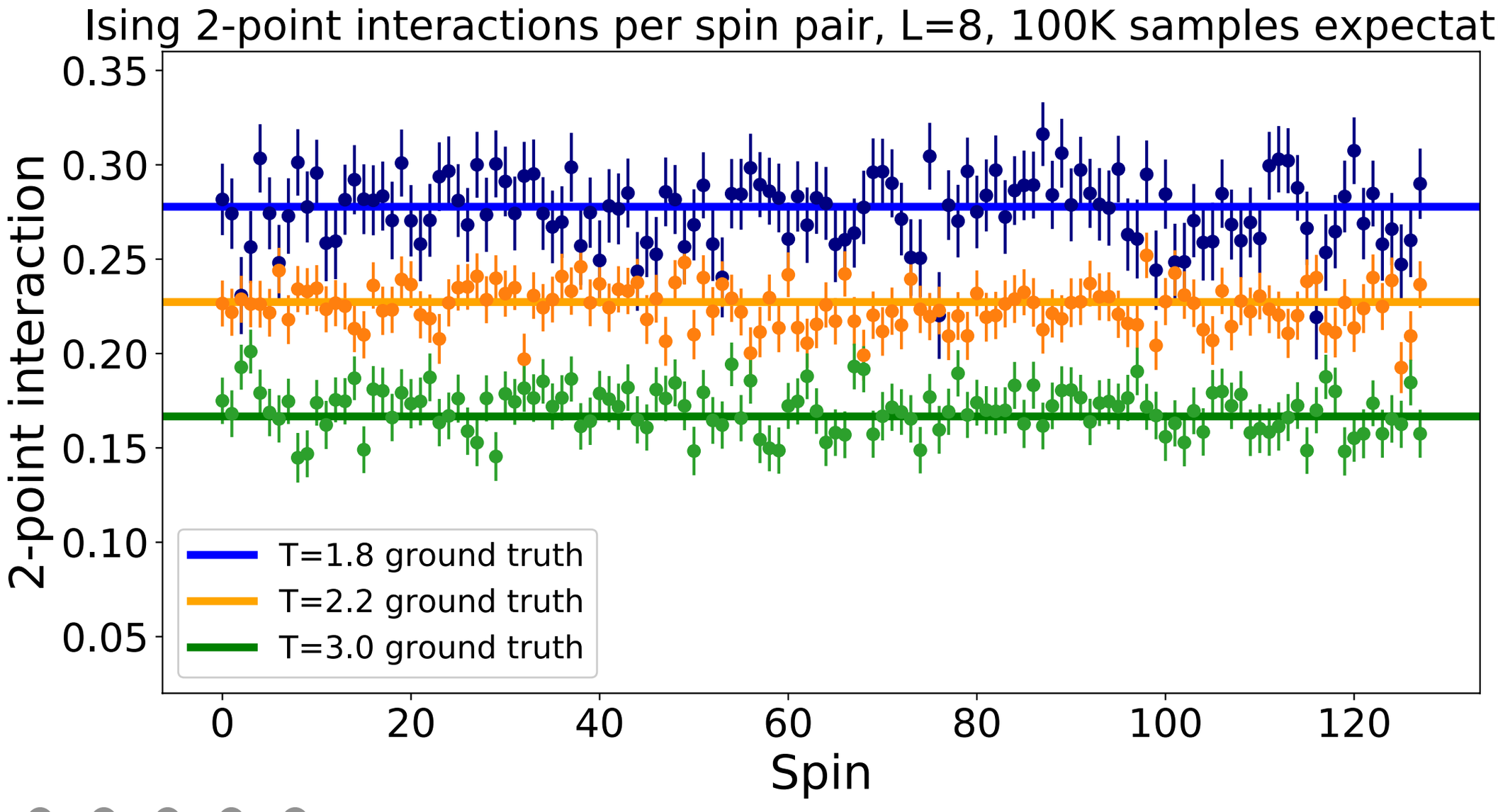
e.g. 3-point interaction: ...

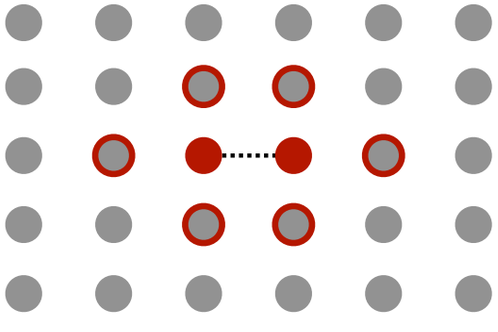
$$\frac{1}{6} \sum_i \ln \frac{(1 + e^{c_i + W_{ij_1} + W_{ij_2} + W_{ij_3}})(1 + e^{c_i + W_{ij_1}})(1 + e^{c_i + W_{ij_2}})(1 + e^{c_i + W_{ij_3}})}{(1 + e^{c_i + W_{ij_1} + W_{ij_2} +})(1 + e^{c_i + W_{ij_1} + W_{ij_3}})(1 + e^{c_i + W_{ij_2} + W_{ij_3}})(1 + e^{c_i})}$$

Closed form expression!

Cossu et. al., Physical Review B (2018)

Back to Ising ...



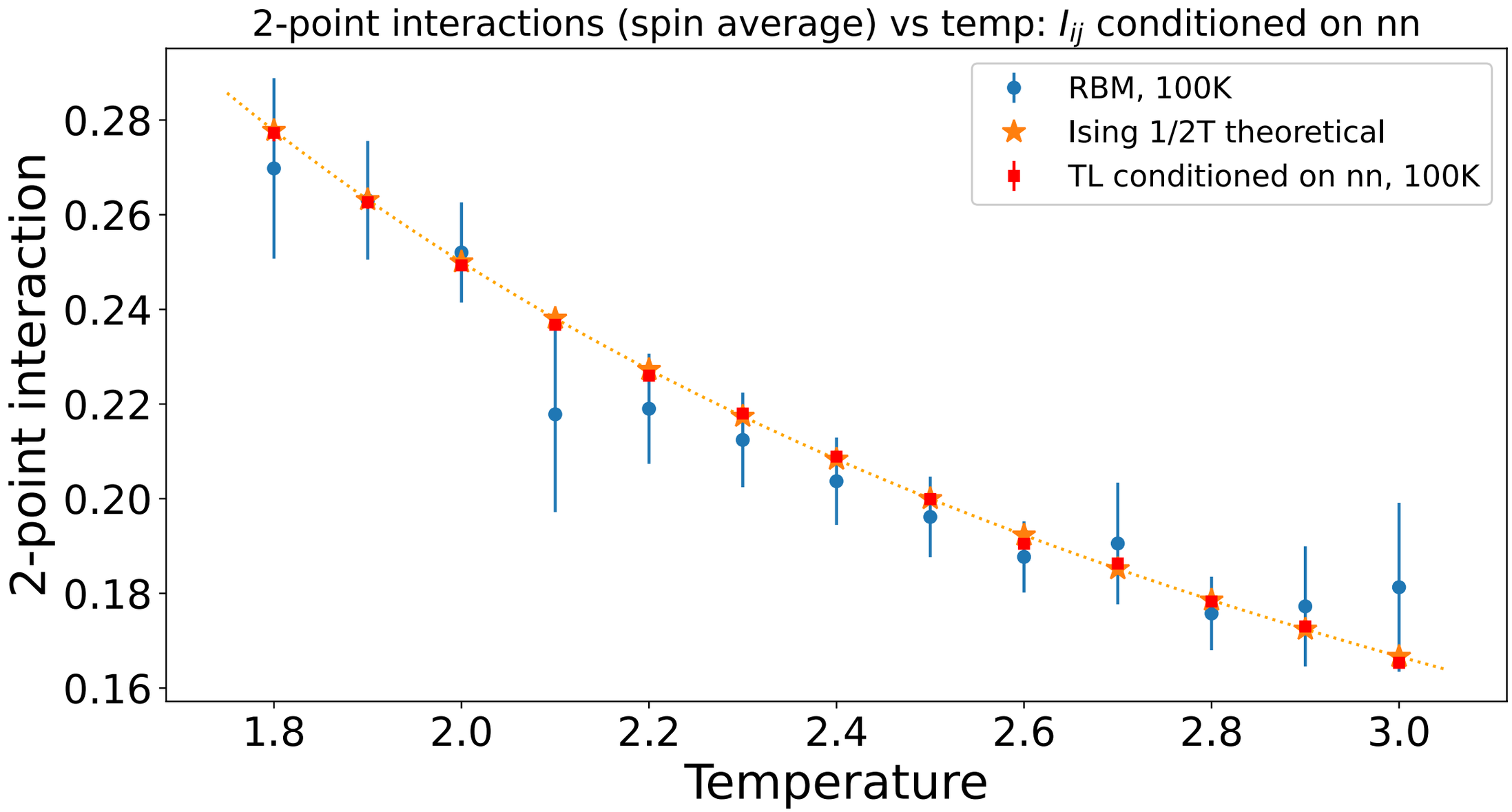


$$H(s) = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j$$

Beentjes & Khamseh, Physical Review E (2020)

Model-independent estimation results

Conditioning on parent spins to isolate pairs from the rest of the system (Markovian). Run time: Few seconds per temperature.



100K samples