Numerical challenges in extracting features from biological data using neural networks

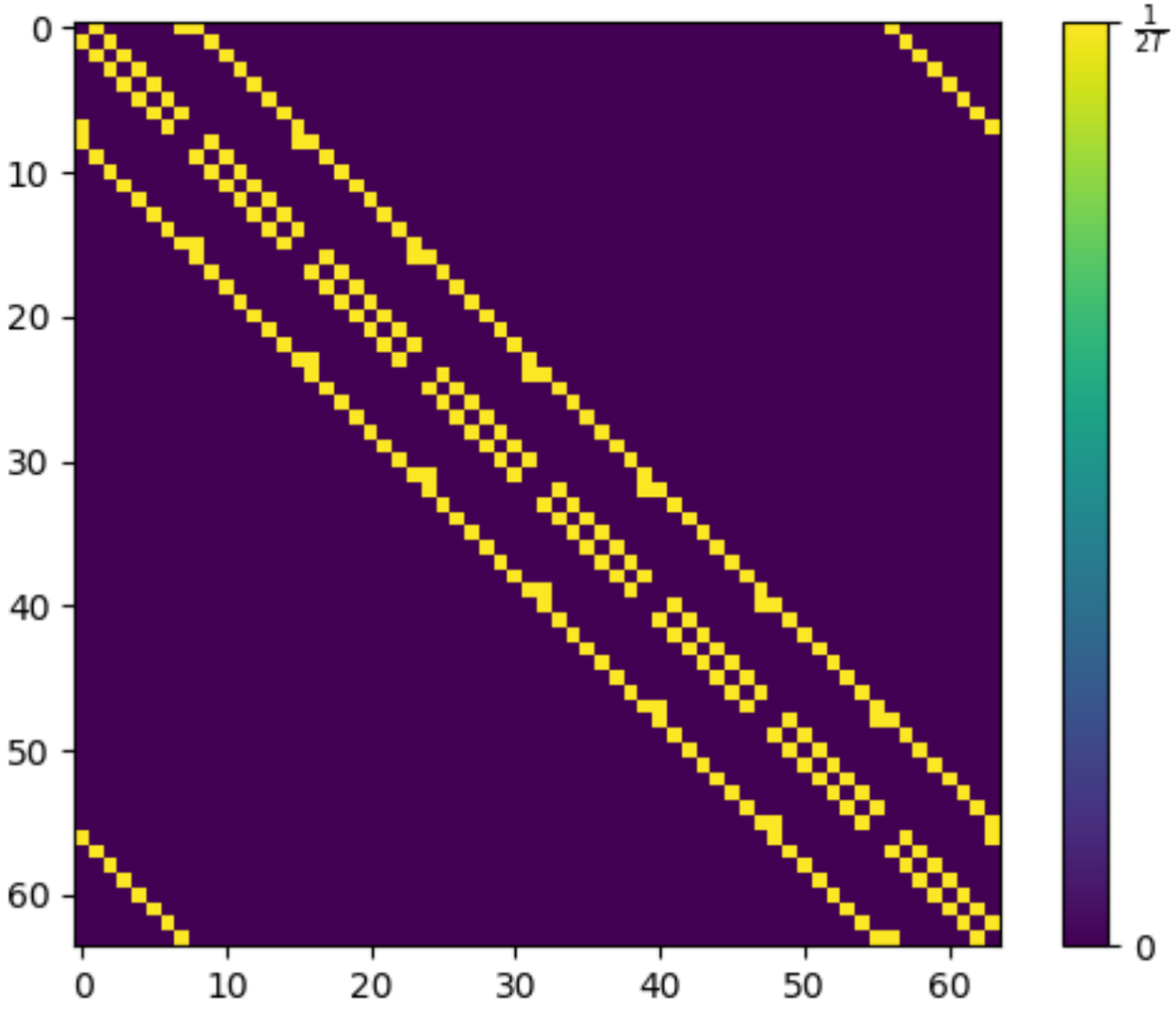
Ava Khamseh

June 19, 2019

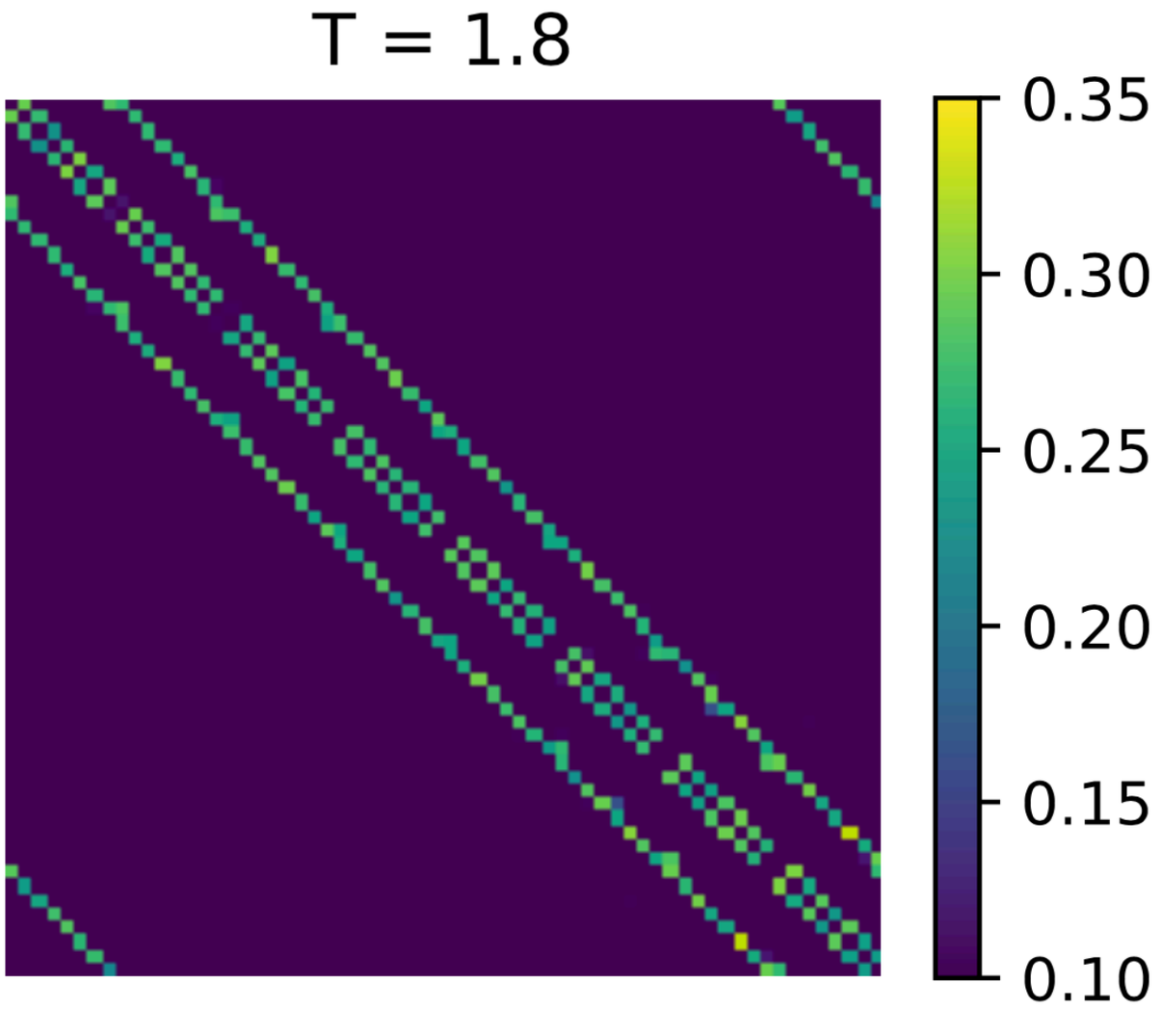
Ascent Robotics Inc: Guido Cossu
IGMM: Abel Jansma, Chris Ponting
Higgs Centre: Luigi Del Debbio,Tommaso Giani, Michael Wilson

RBM Predictions: Couplings

Ising Model



RBM prediction



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

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

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

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

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

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.

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:

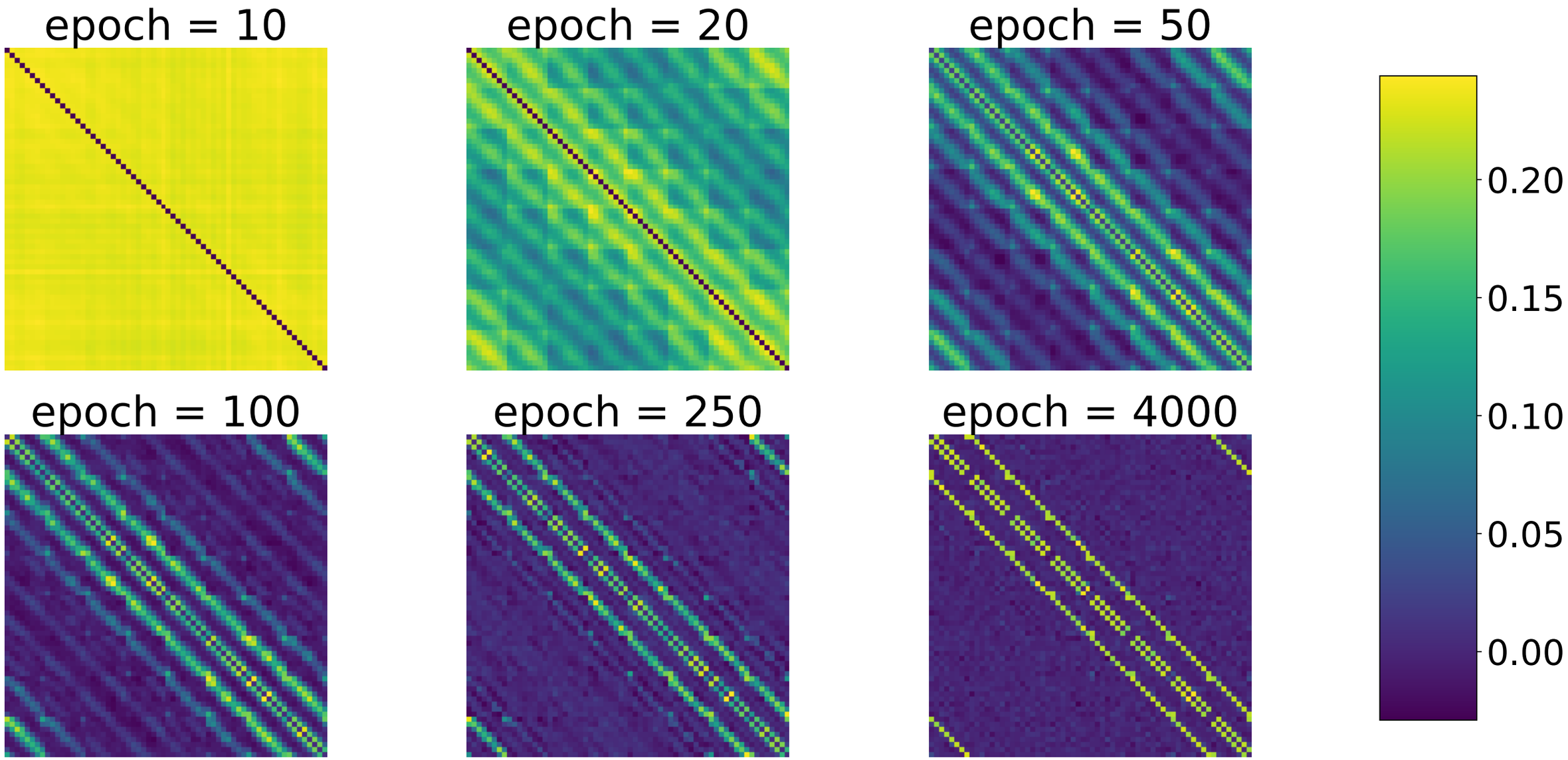
$$\sum_{n>1} \frac{1}{2(n!)} \sum_{0<k<n} \sum_{j_1 \neq j_2} \left(\sum_i \kappa_i^{(n)} \binom{n}{k} W_{ij_1}^k W_{ij_2}^{n-k} \right) v_{j_1} v_{j_2}$$

...

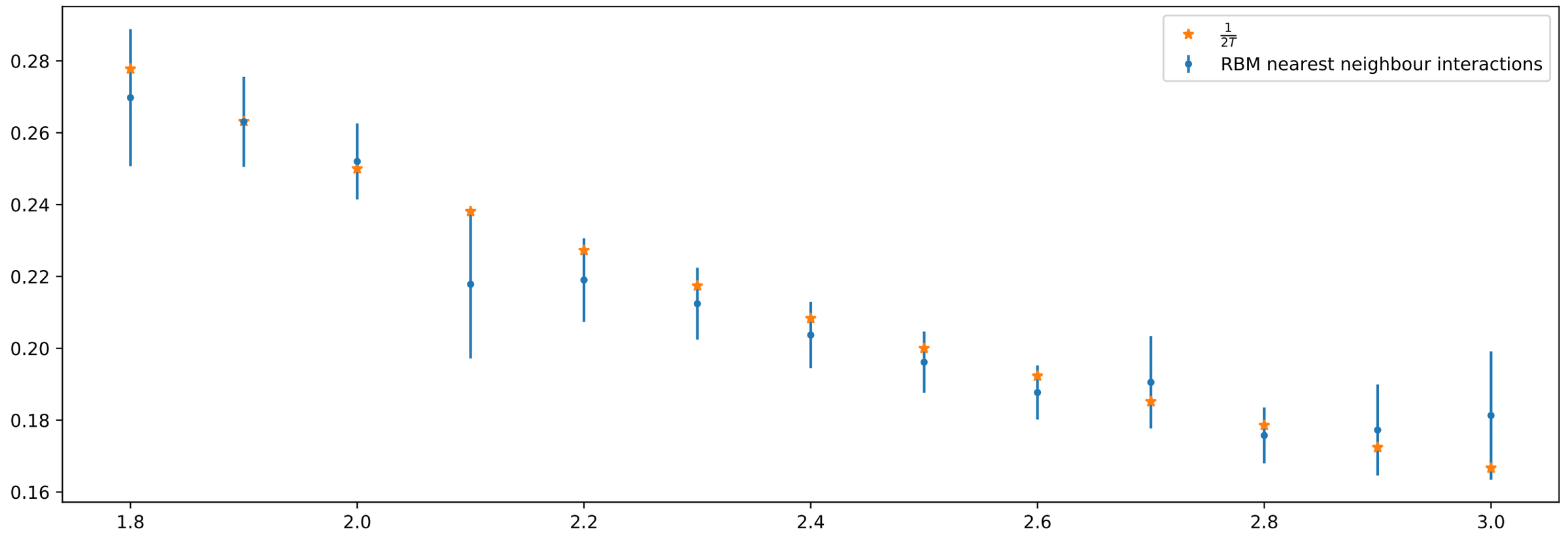
$$H_{j_1 j_2} = \frac{1}{8} \sum_i \ln \frac{(1 + e^{c_i + W_{ij_1} + W_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + W_{ij_1}})(1 + e^{c_i + W_{ij_2}})}$$

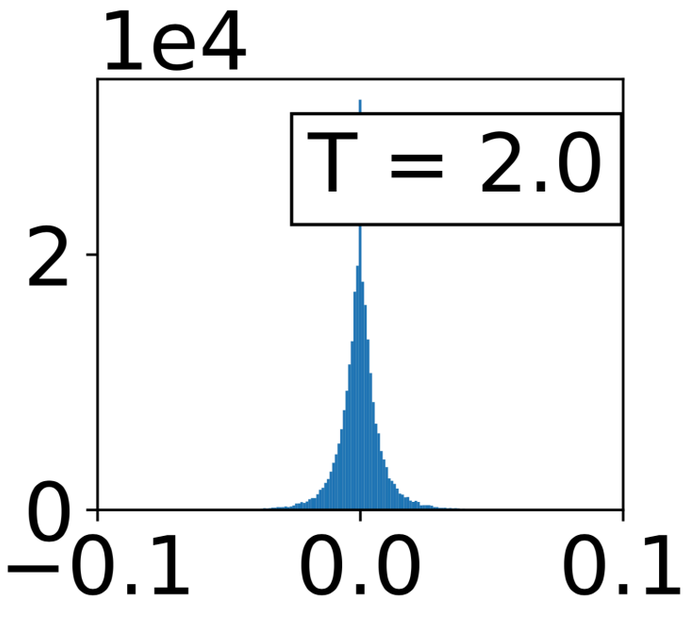
Closed form expression!

Couplings during training

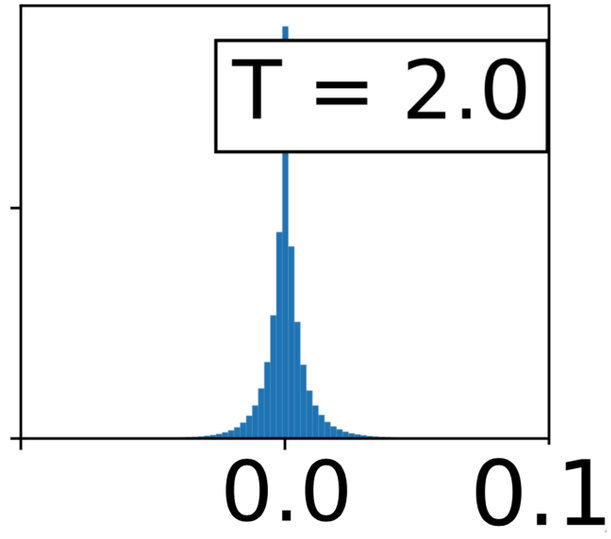


RBM Predictions: Couplings

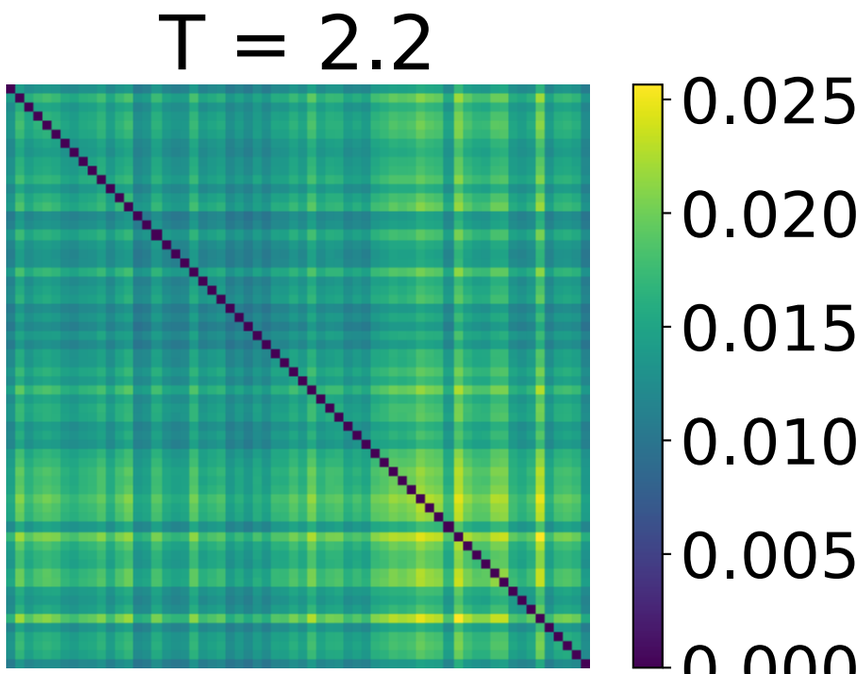




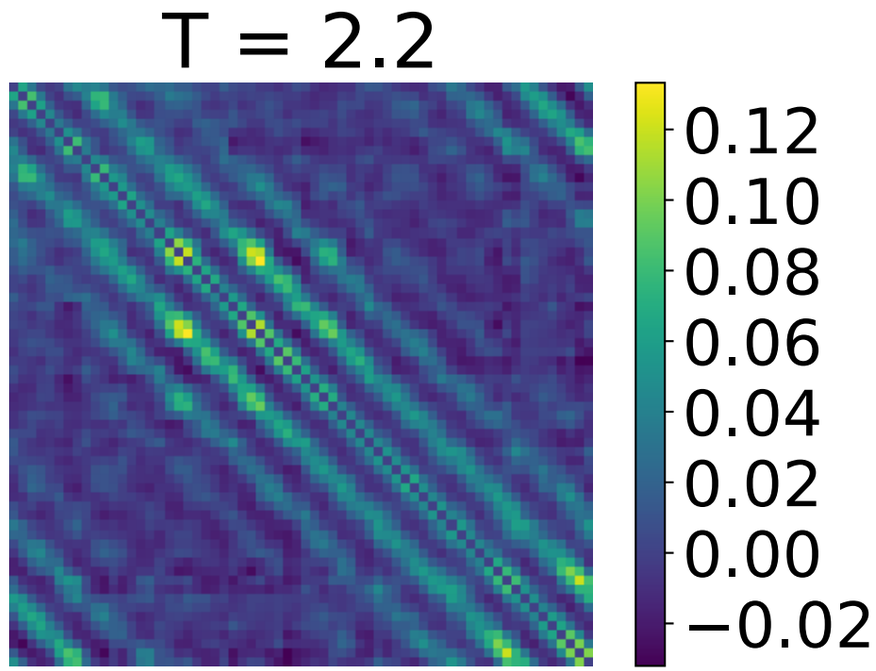
3- and 4-point couplings



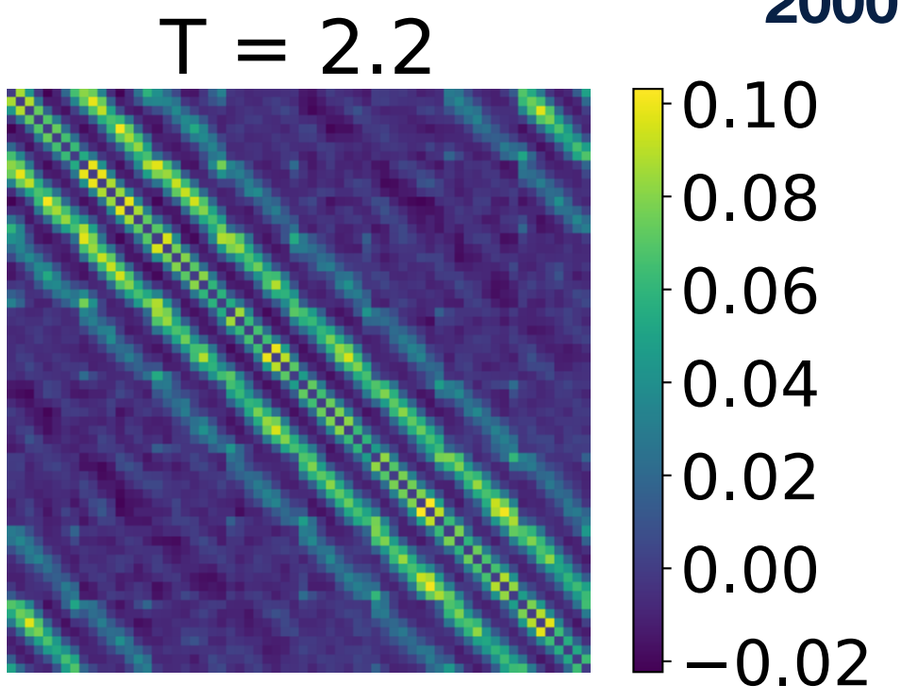
Small number of training examples



200 Examples



2000 Examples



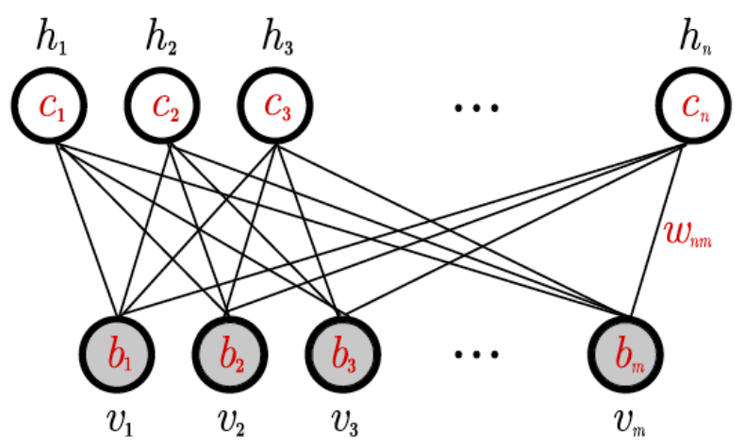
10000 Examples

RBM for the UK Biobank

- Training examples: 452264 patients from GeneAtlas

- Each visible node is a binary (e.g. disease) or binarised trait

- 67 traits in total

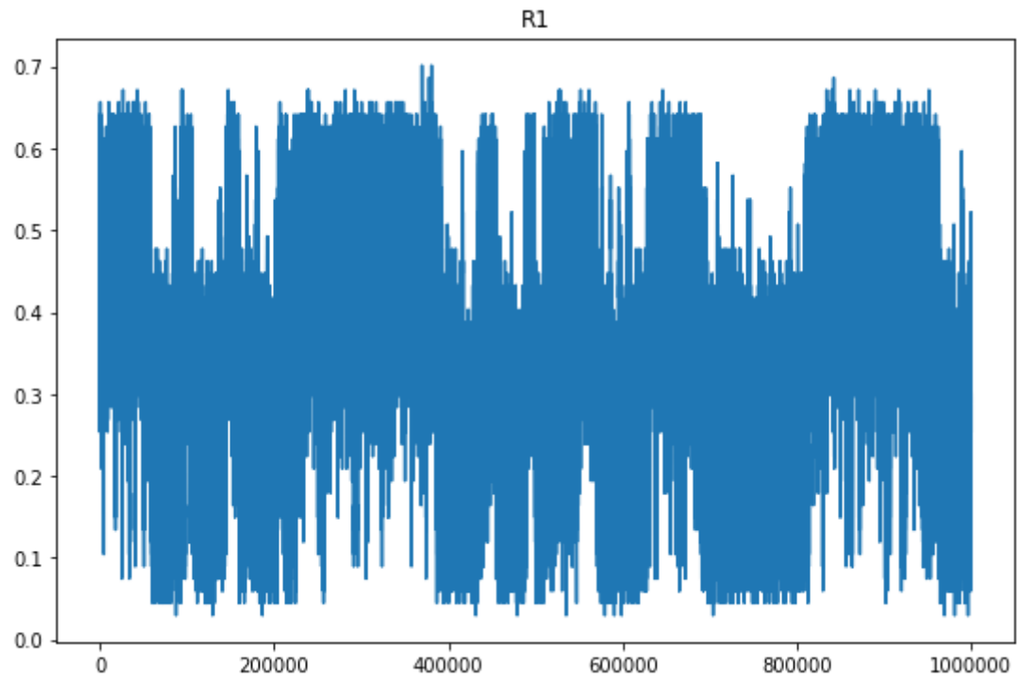


- Contains **controls** such as tea-intake and number of cars

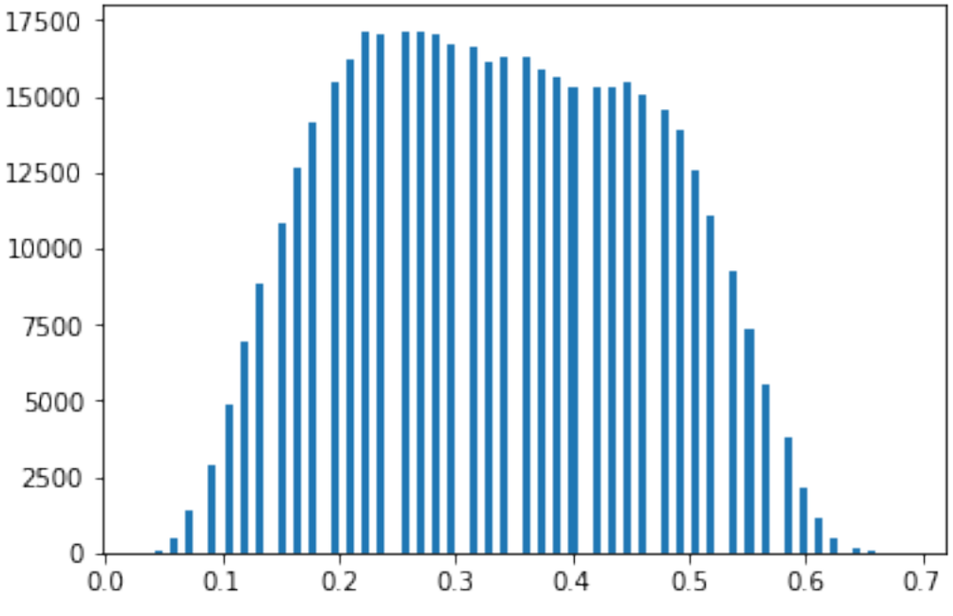
- Aim: Extract **all-to-all** coupling strengths between traits

RBM for the UK Biobank: Preliminary Results

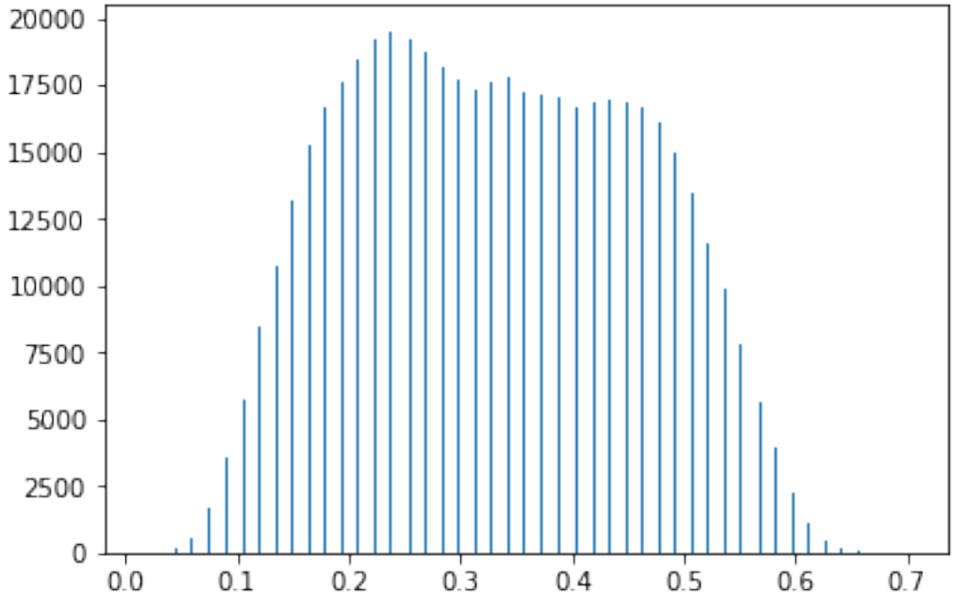
- Generate patients from the trained RBM
- Follow the history of the first moment of the distribution



Raw Data

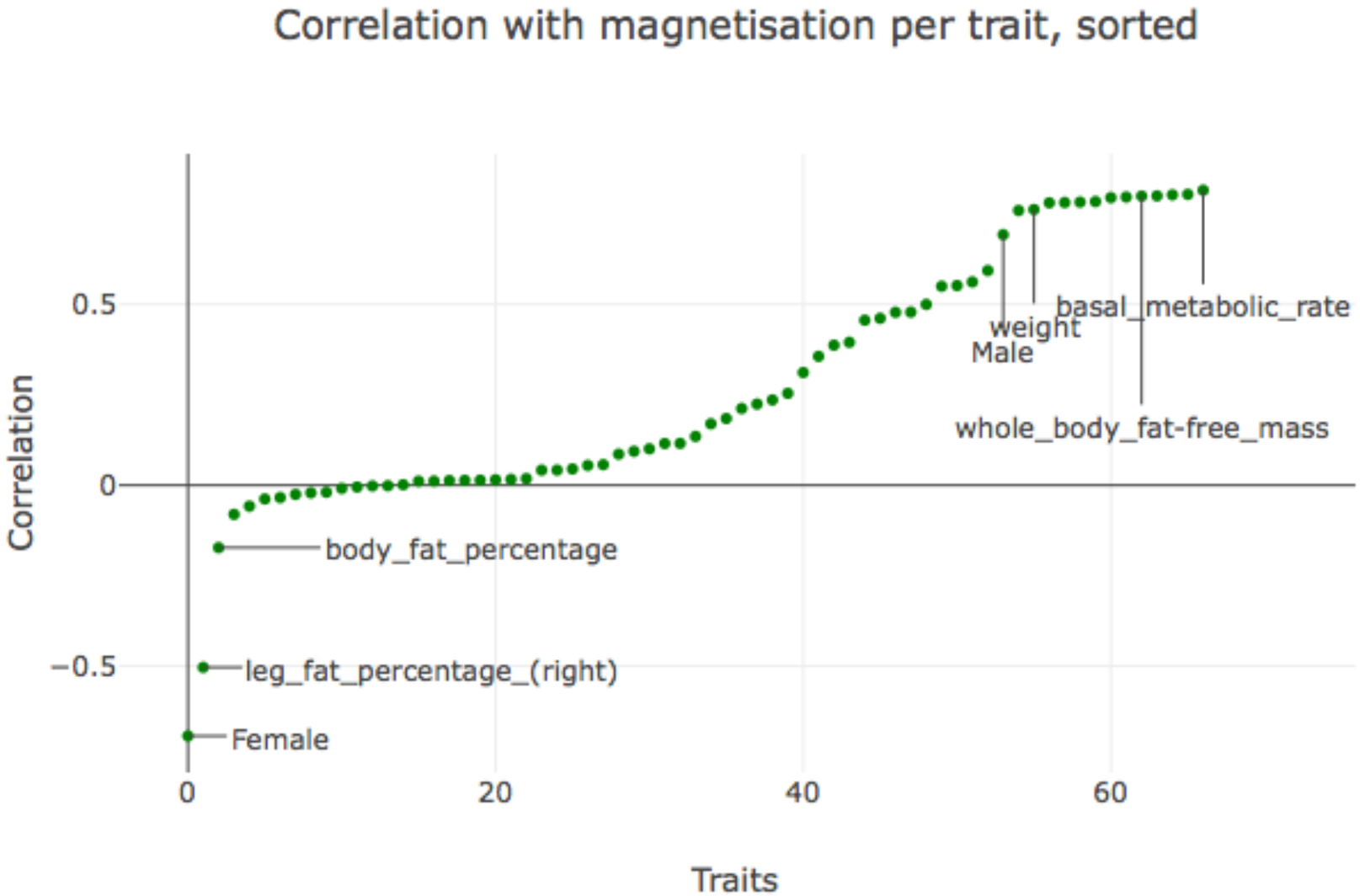


Generated Data



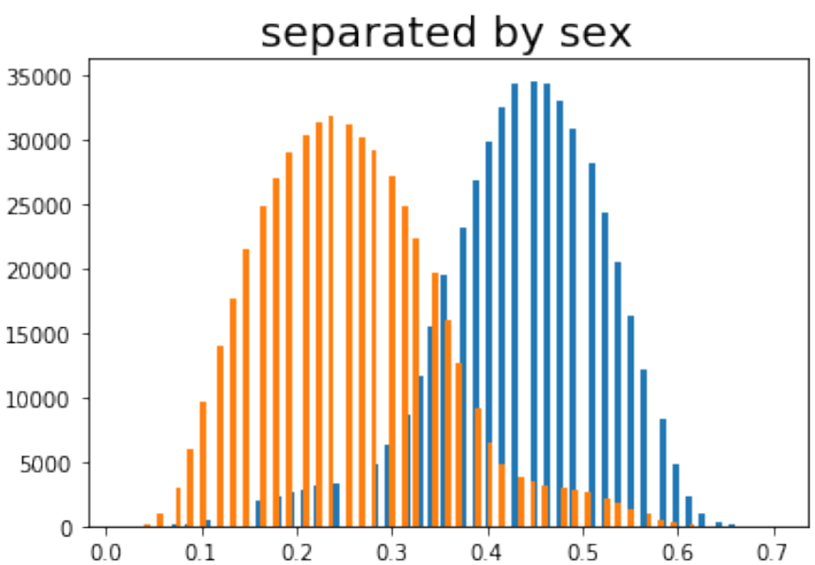
RBM for the UK Biobank: Preliminary Results

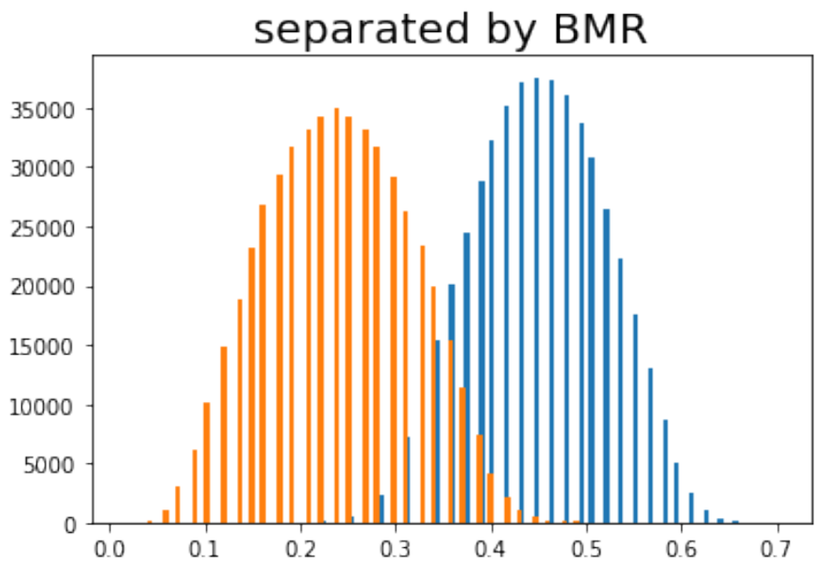
- Two main states: Large/small first moment
- Male vs Female, high vs low metabolic rate, low vs high fat mass!

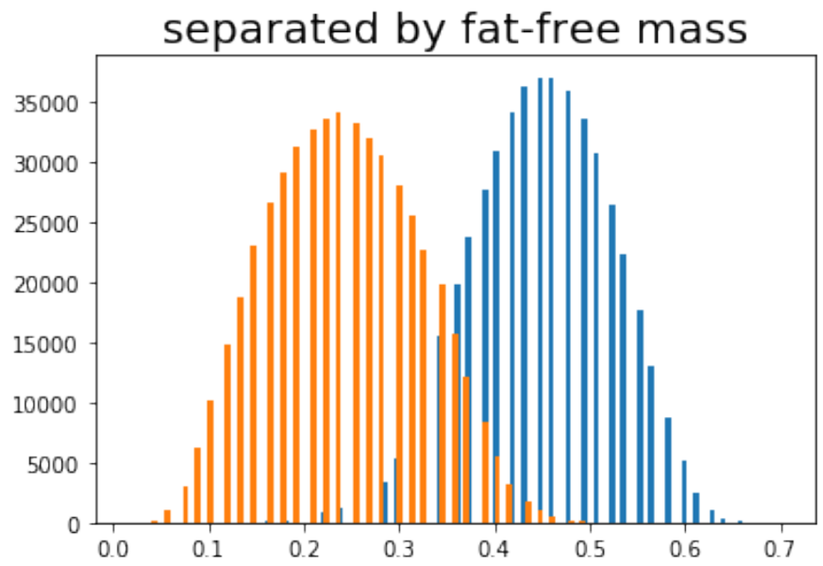


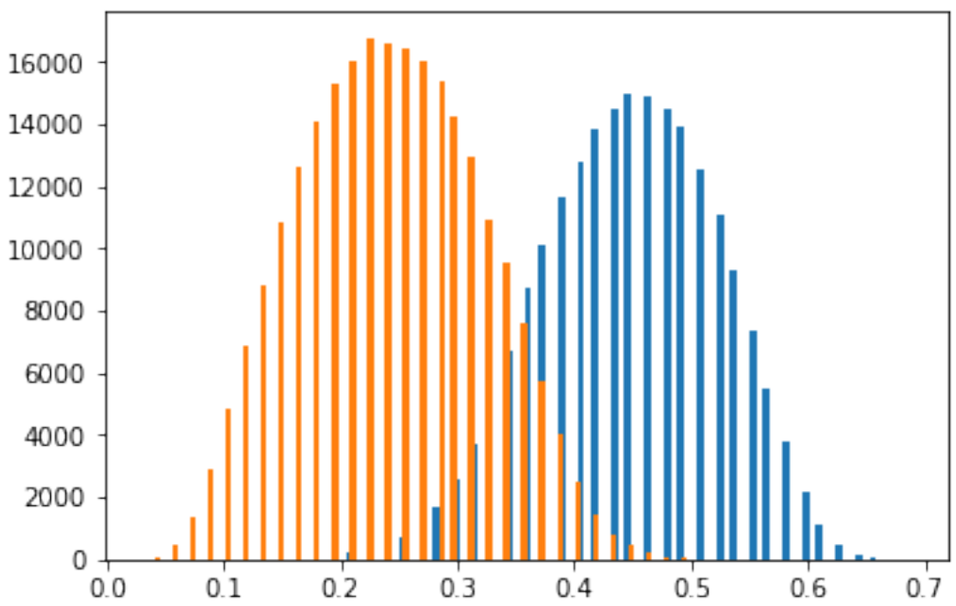
RBM for the UK Biobank: Preliminary Results

Generated by the RBM:

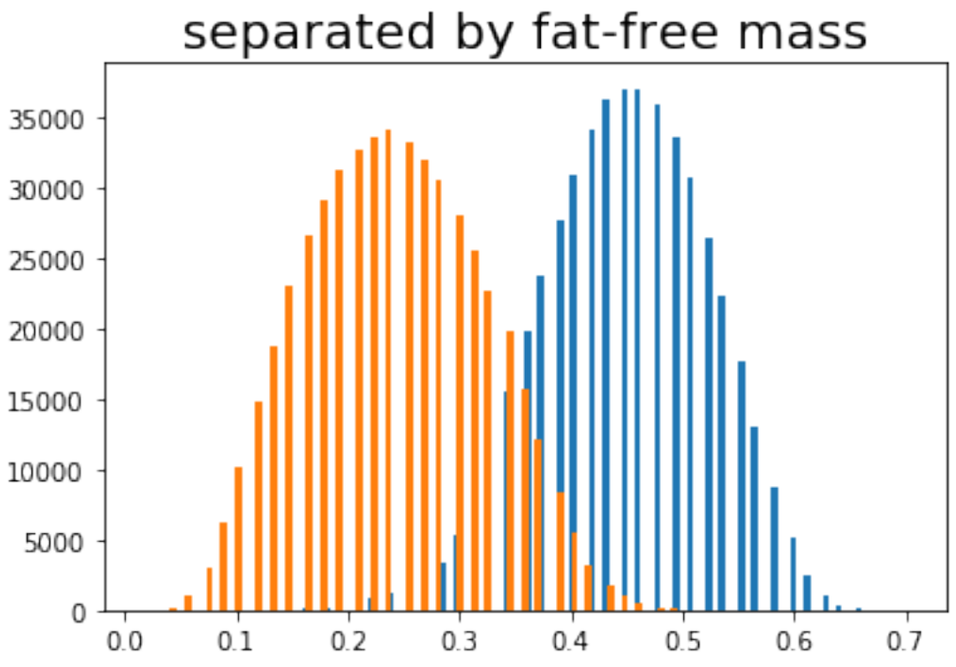






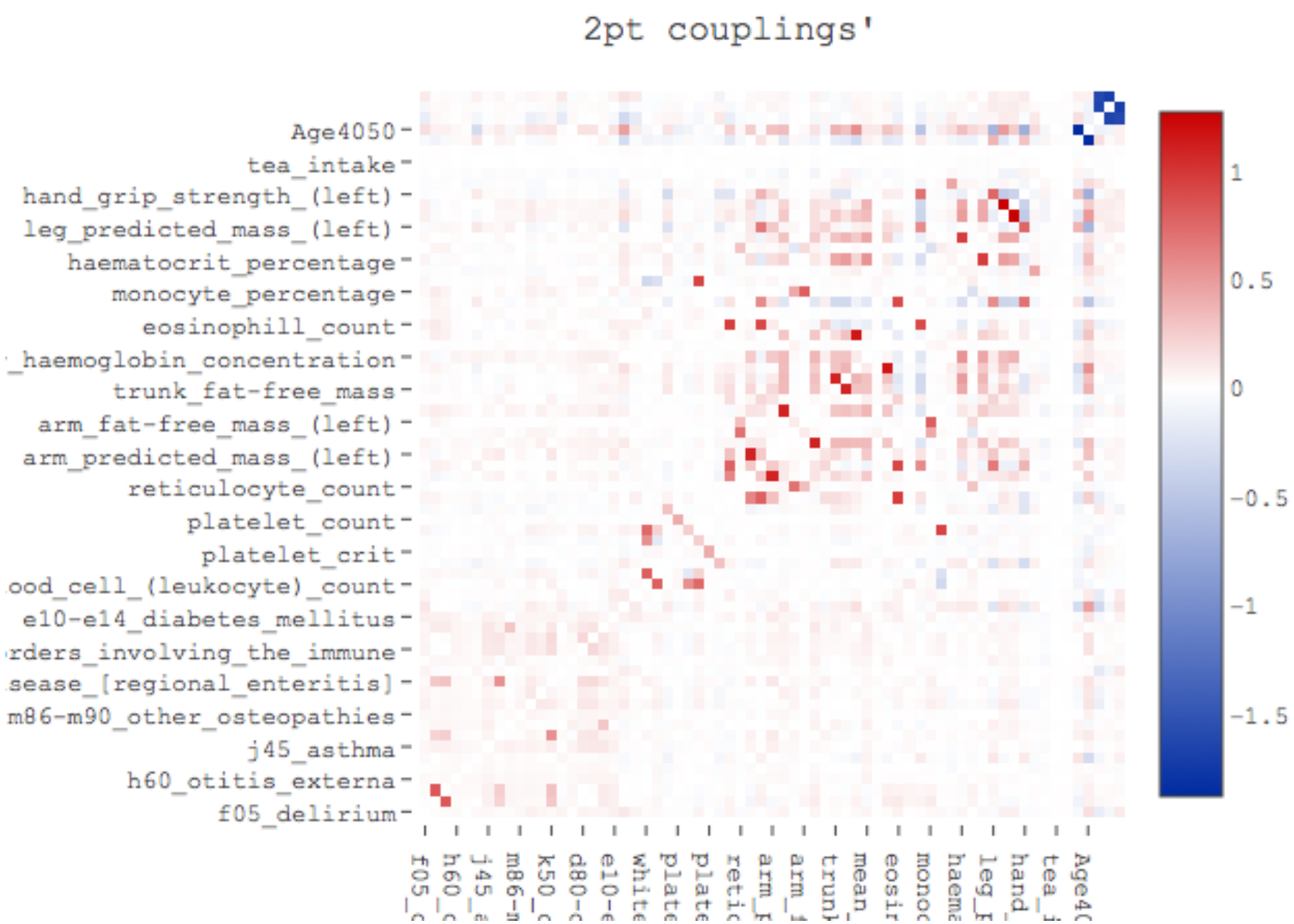


Raw Data



Generated Data

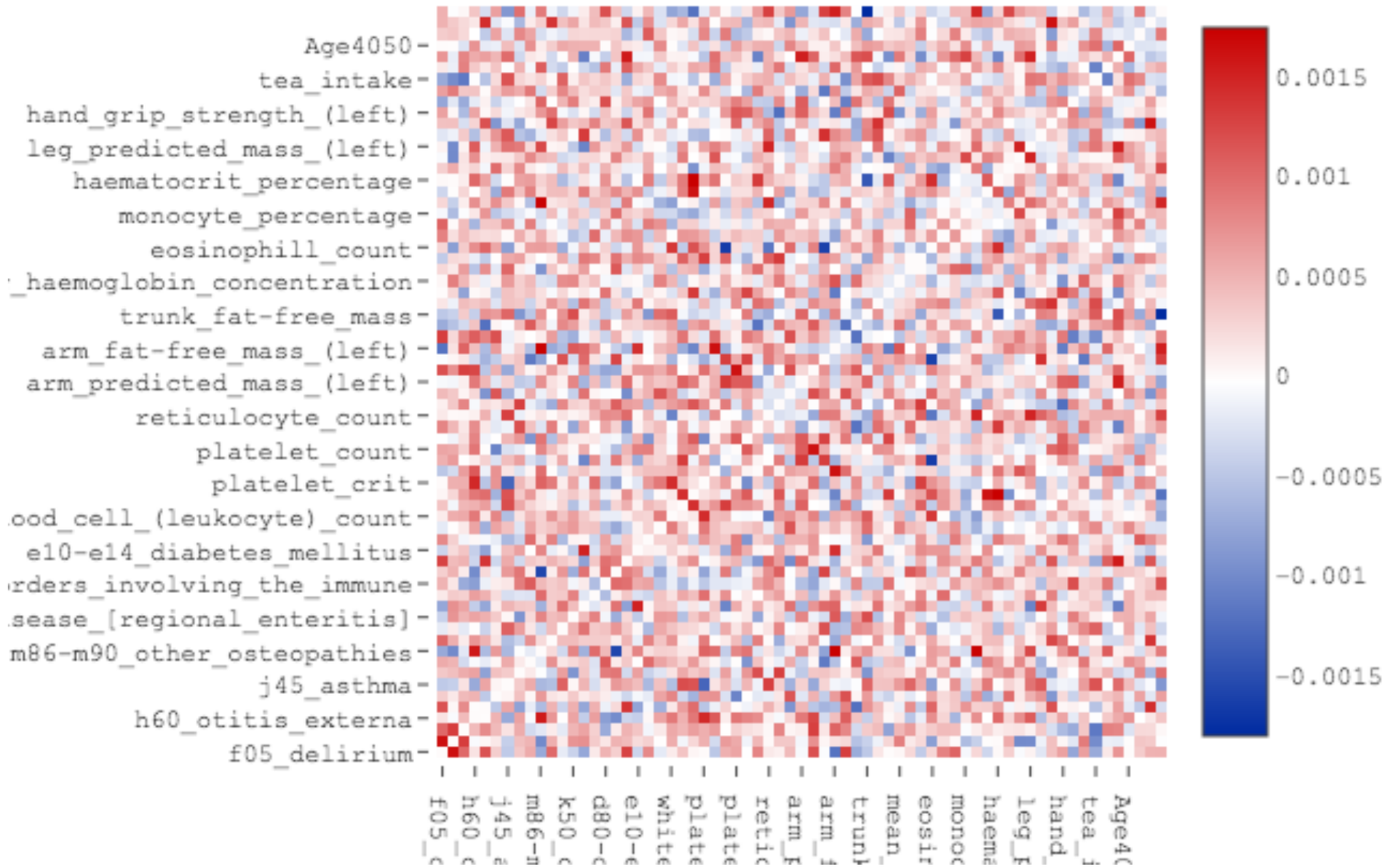
RBM for the UK Biobank: Coupling Matrix



- Different as compared to correlation directly measured from data?
- Error bars?

Training: Revisited

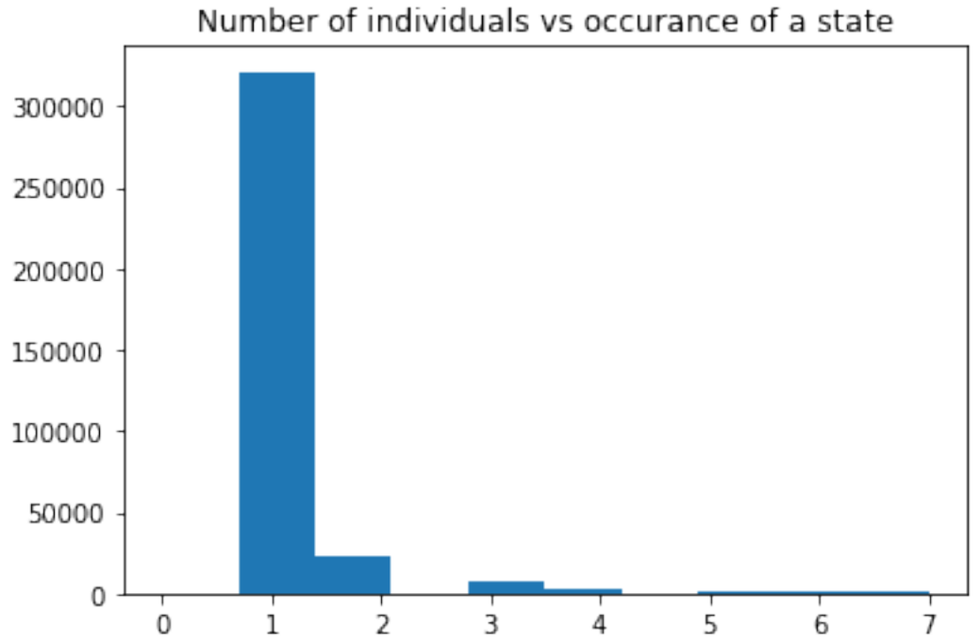
- Statistical error due to fluctuations during training
- Random permutations, very small error margin

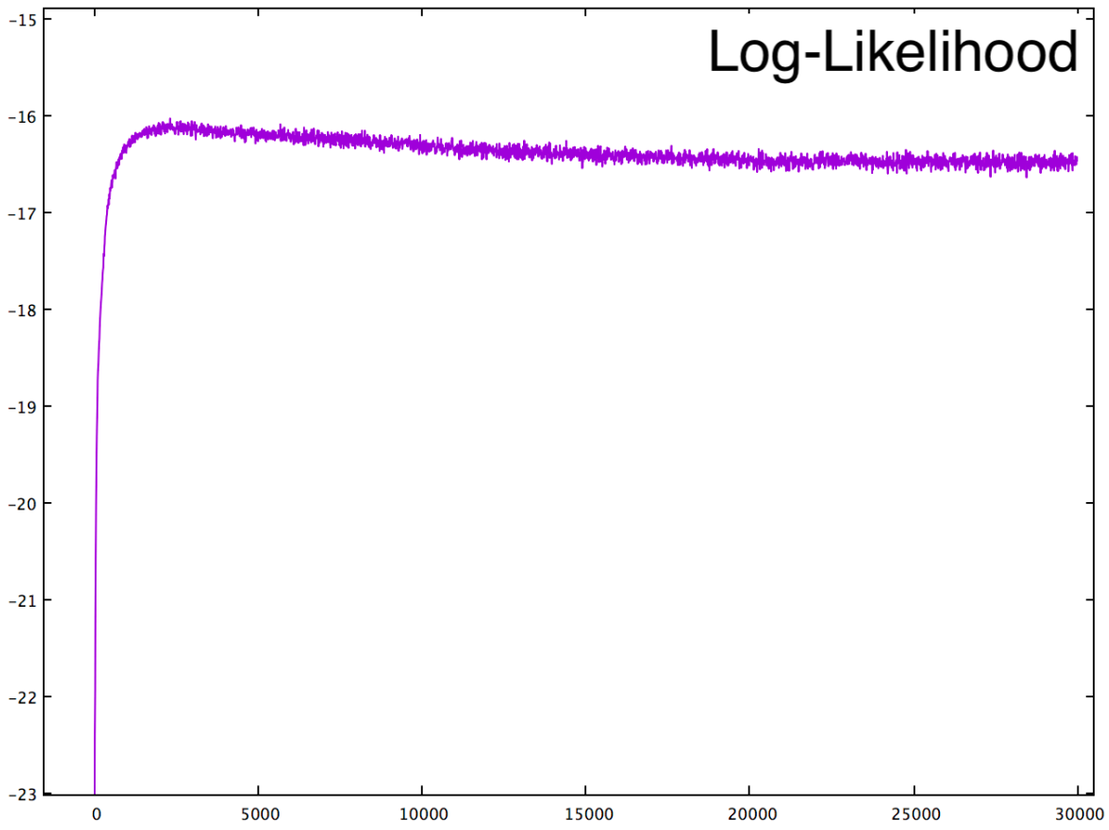


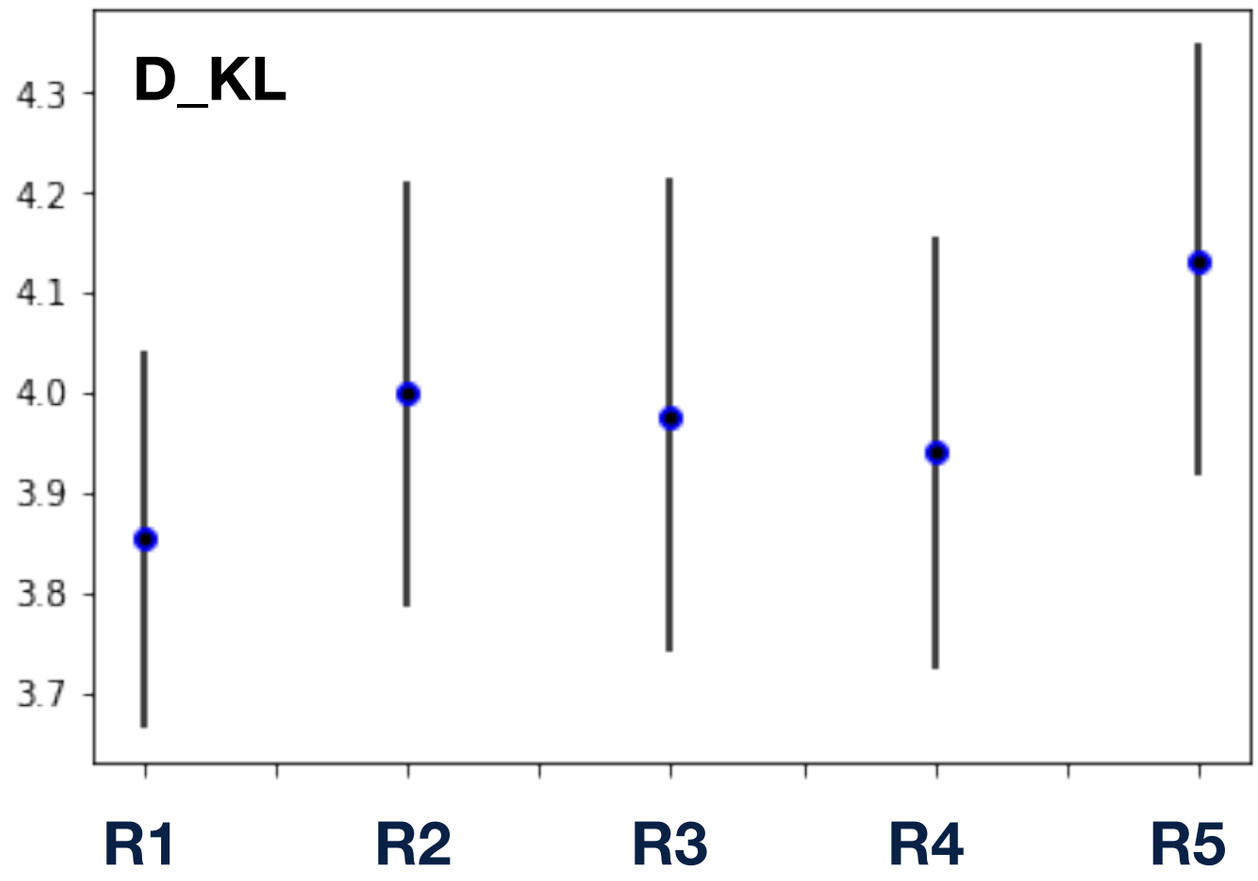
Random initialisation on 5 RBMs: KL Divergence

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

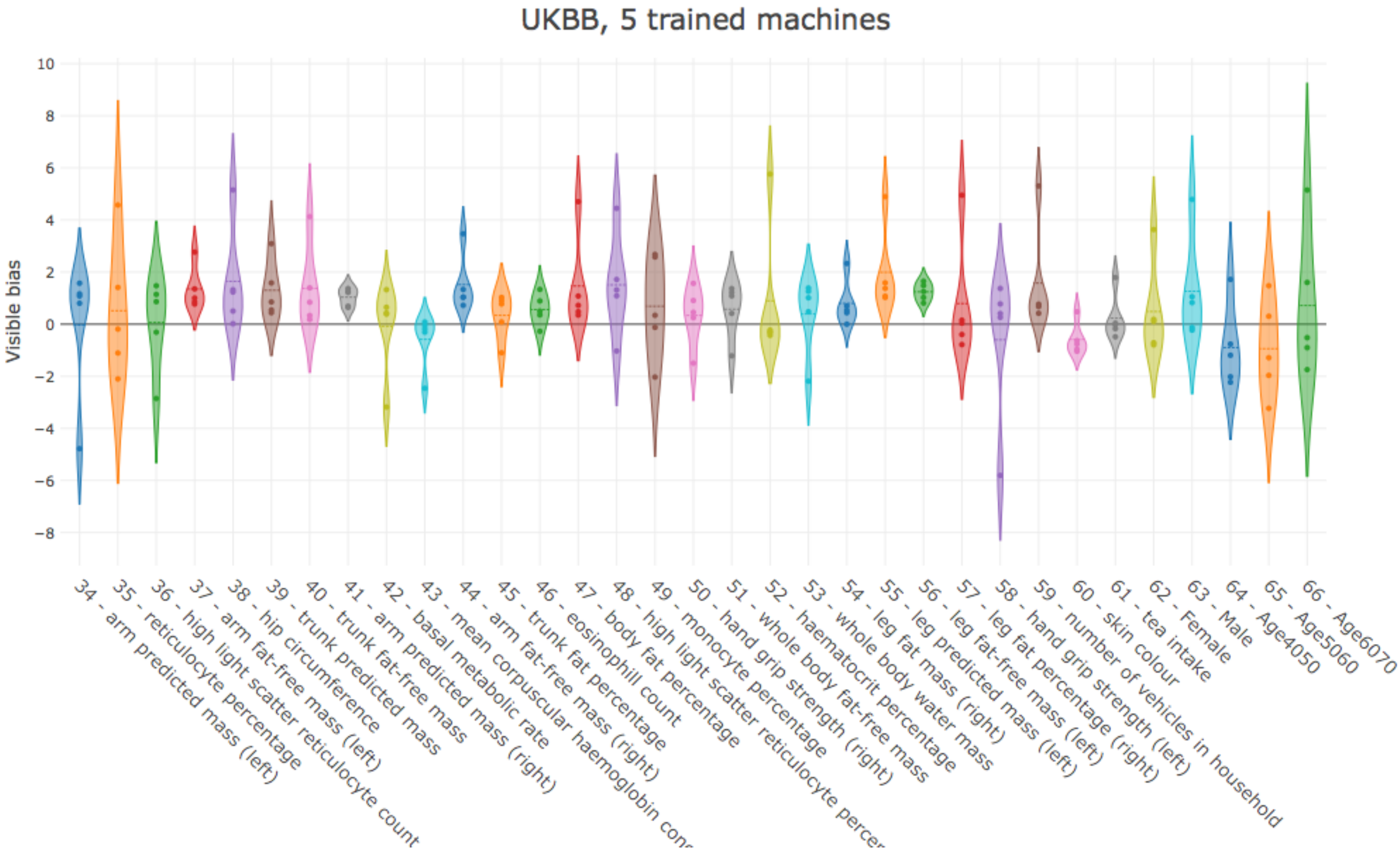
$$= \sum_{\mathbf{v}} \left(q_{\mathrm{data}}(\mathbf{v}) \log\left(q_{\mathrm{data}}(\mathbf{v})\right) - q_{\mathrm{data}}(\mathbf{v}) \log\left(p_\theta(\mathbf{v})\right) \right)$$



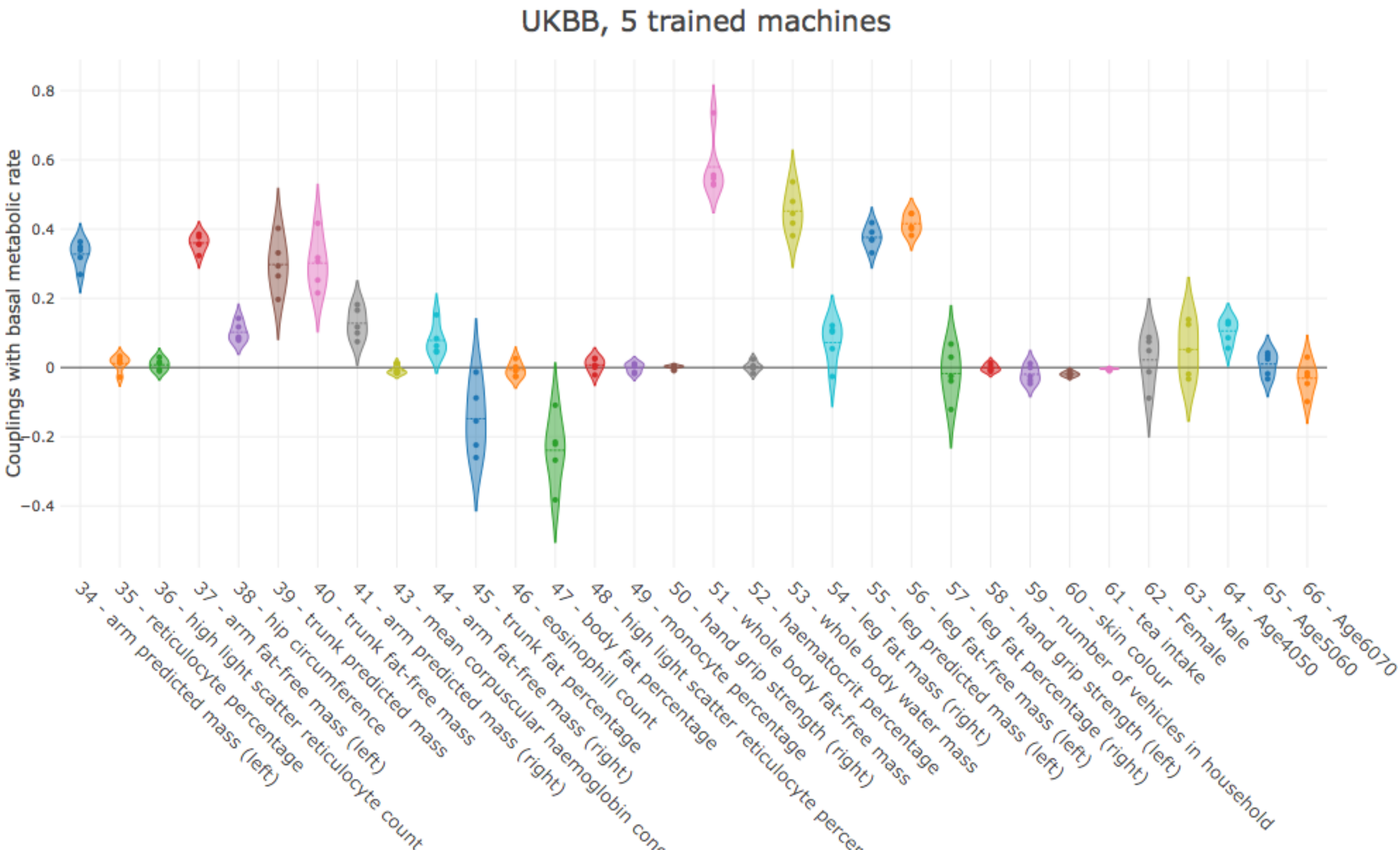




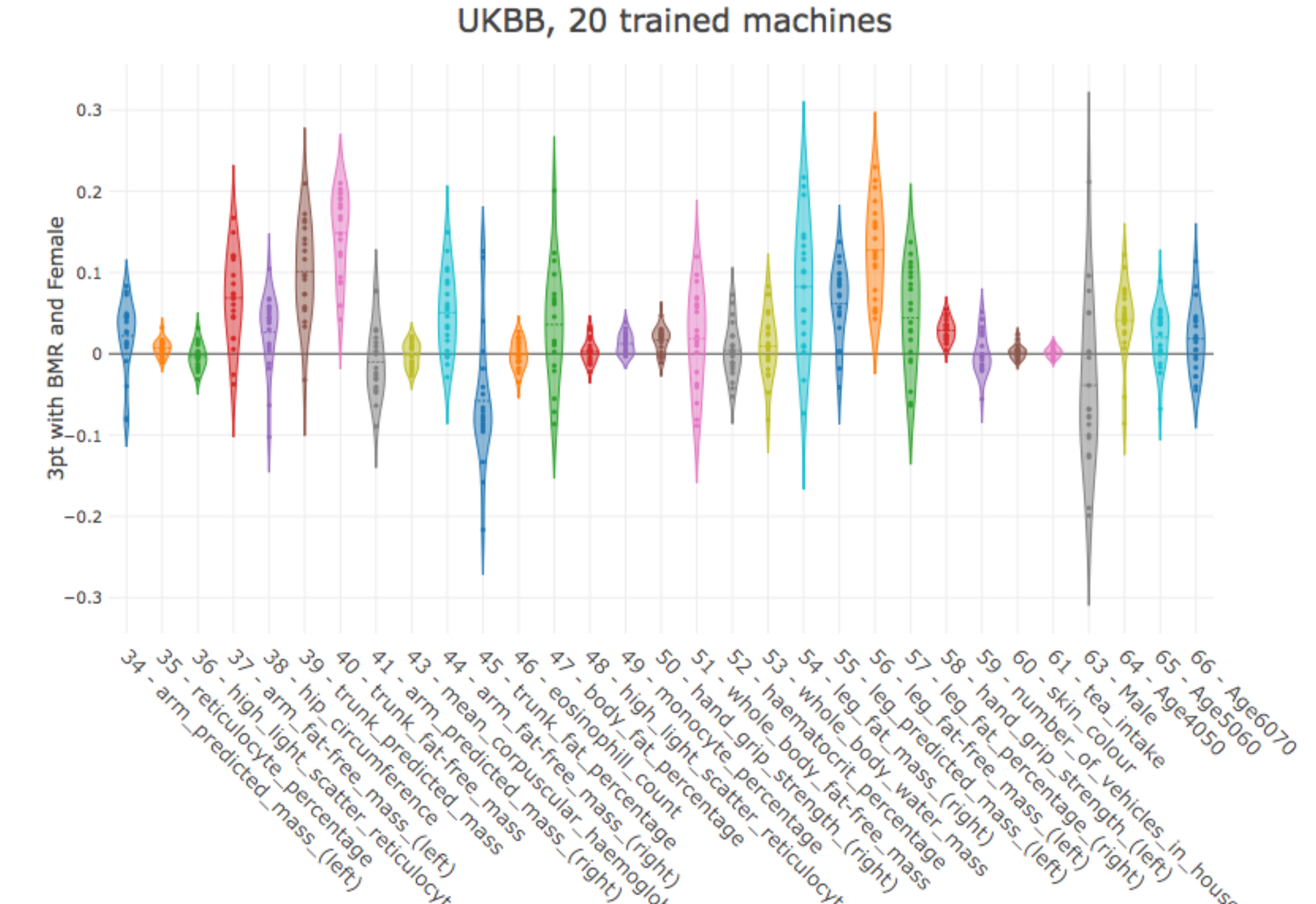
Bias terms across several RBMs



RBM for the UK Biobank: Coupling across 5 RBMs



RBM for the UK Biobank: Coupling across 20 RBMs



Work in progress

- Identifying relevant biological signals from 2-point couplings
- 3-point couplings, successful for an ‘Ising’-like system
- Derive meaning from 3-point couplings for biology (absence of a Hamiltonian)
- Analytical computations and training on smaller RBMs

Numerical challenges in extracting features from biological data using neural networks

Ava Khamseh

June 19, 2019

Ascent Robotics Inc: Guido Cossu
IGMM: Abel Jansma, Chris Ponting
Higgs Centre: Luigi Del Debbio,Tommaso Giani, Michael Wilson