Data modeling with Restricted Boltzmann Machines

Decelle, Furtlehner, Navas Gómez, Seoane,
SciPost Phys (2024) arXiv:2309.02292
& follow-up

Beatriz Seoane
LISN Paris-Saclay University

Inferring effective couplings with Restricted Boltzmann Machines

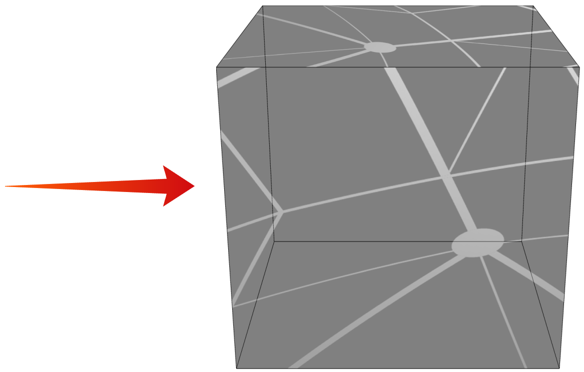
Beatriz Seoane

LISN Paris-Saclay University

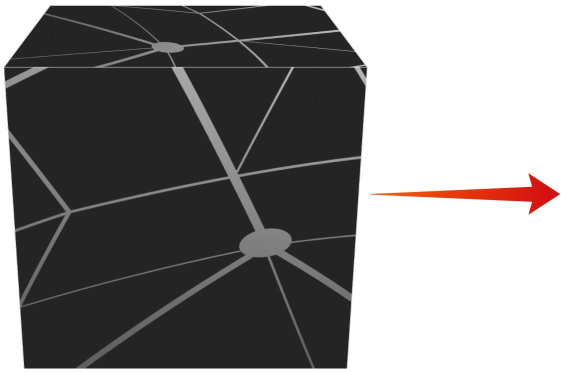
Decelle, Furtlehner, **Navas Gómez**, Seoane,
SciPost Phys (2024) arXiv:2309.02292
& follow-up

Introduction : Generative approach

training

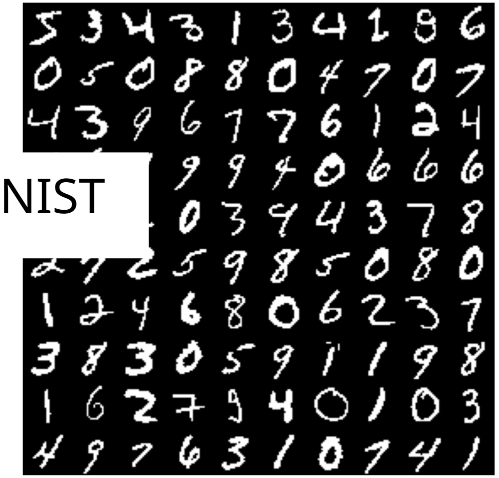


generating

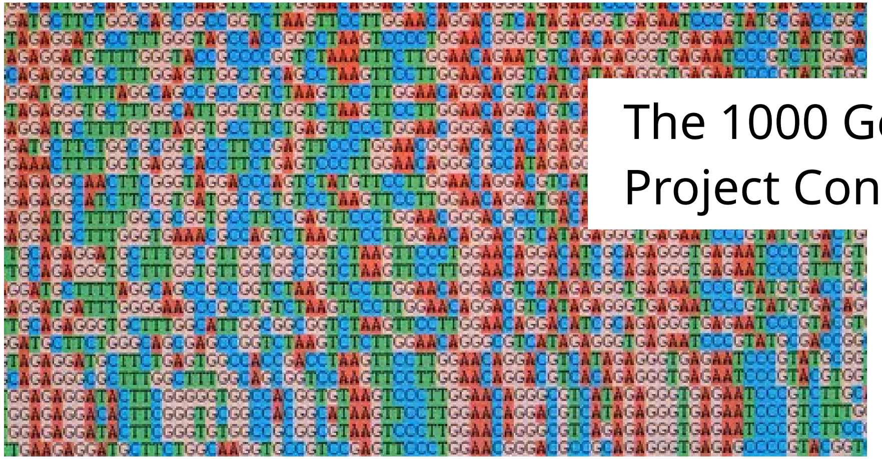


- **Energy based models (RBMs, Generative Convnets)**
- **Diffusion models,** normalizing flows
- Variational AutoEncoder (VAE)
- Generative Adverarial Network (GAN)
- Autoregressive methods

Introduction : generative approach

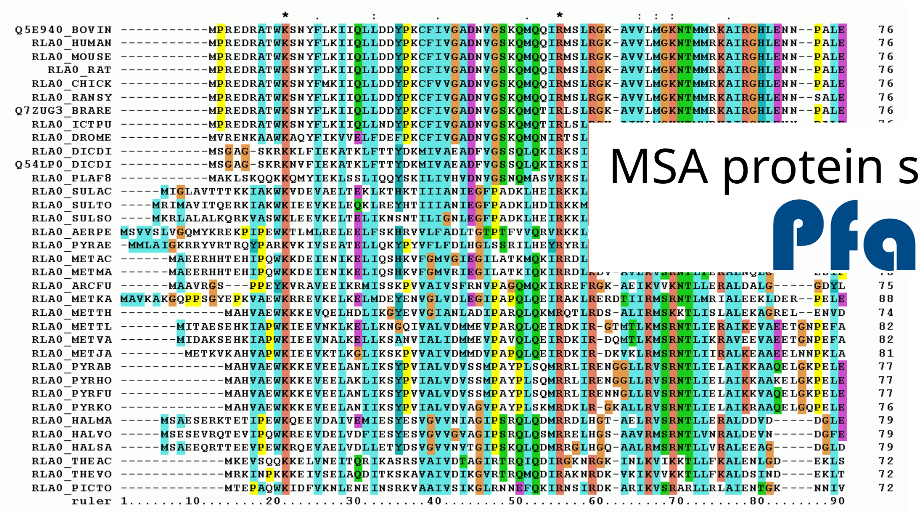


MNIST



The 1000 Genomes
Project Consortium

CELEBA



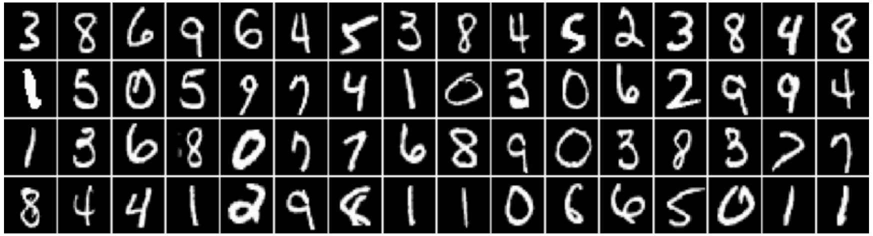
MSA protein sequences

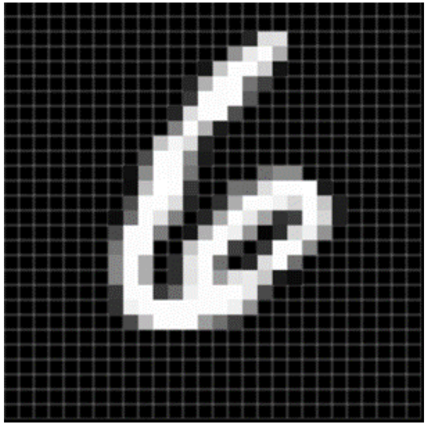
Pfam

Energy based models (EBMs)

Hinton, Hopfield, LeCun, Bengio

- Dataset $X = \left\{ \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)} \right\}$





Empirical *Model*

$$p_{\text{data}}(x) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Boltzmann distribution

$$E_{\boldsymbol{\theta}}(\boldsymbol{x})$$

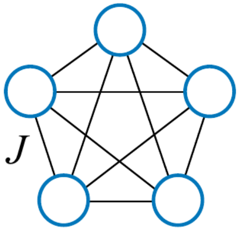
Learning : adjust the parameters so that the dataset configurations are typical configurations of the model.

Energy based models (EBMs)

visible variables

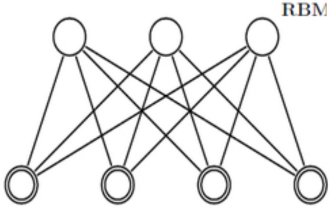
$$\mathcal{E}(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$$

Latent variables



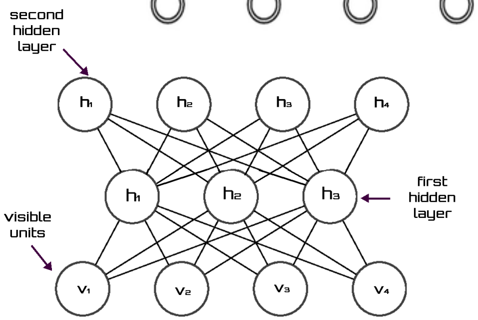
- Boltzmann Machines (Ising/Hopfield/Potts models)

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.



- Restricted Boltzmann Machines

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



- Deep Boltzmann Machines

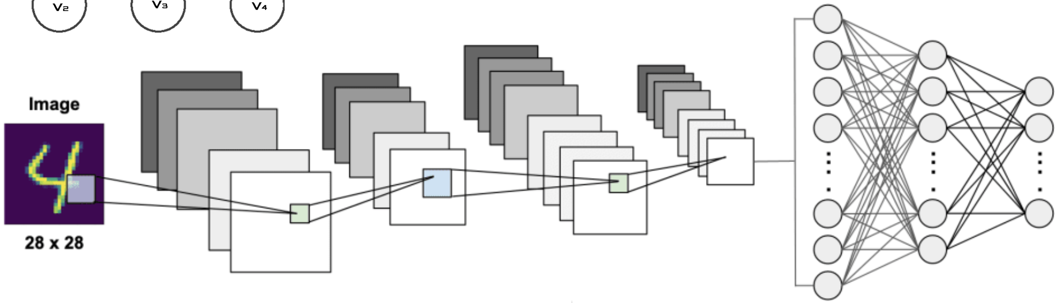
-Ruslan Salakhutdinov, Geoffrey Hinton (2009) Deep Boltzmann Machines.

-Bengio, Y. (2009). *Learning deep architectures for AI.*

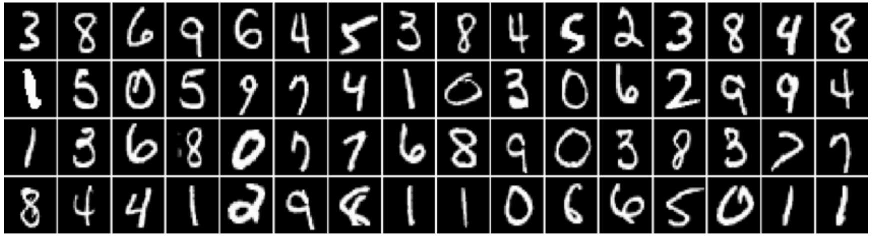
- Generative ConvNets

- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). *A tutorial on energy-based learning.*

- Xie, J., Lu, Y., Zhu, S. C., & Wu, Y. (2016, June). *A theory of generative convnet.*



Review of the training procedure



Dataset

$$X = \left\{ x^{(1)}, \ldots, x^{(M)} \right\}$$

Empirical *Model*

Goal of the training:

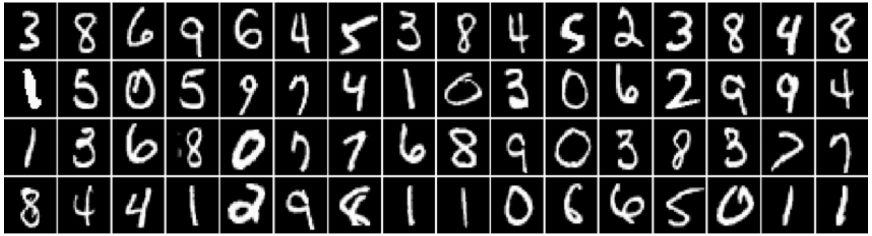
$$p_{\text{data}}(x) \sim p_{\boldsymbol{\theta}}(x) = \frac{e^{-E_{\boldsymbol{\theta}}(x)}}{Z}$$

Minimize Kullback-Leibler (KL) divergence

log-likelihood

$$\begin{aligned} D_{\text{KL}}(p_{\text{data}} || p_{\boldsymbol{\theta}}) &= \sum_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} \\ &= \sum_{\boldsymbol{x}} \underbrace{p_{\text{data}}(\boldsymbol{x}) \log p_{\text{data}}(\boldsymbol{x})}_{\text{Constant}} - \sum_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \end{aligned}$$

Review of the training procedure



Dataset $X = \left\{ x^{(1)}, \ldots, x^{(M)} \right\}$

Goal of the training:

Empirical *Model*

$$p_{\mathrm{data}}(x) \sim p_{\boldsymbol{\theta}}(x) = \frac{e^{-E_{\boldsymbol{\theta}}(x)}}{Z}$$

Maximize the log-likelihood (LL)

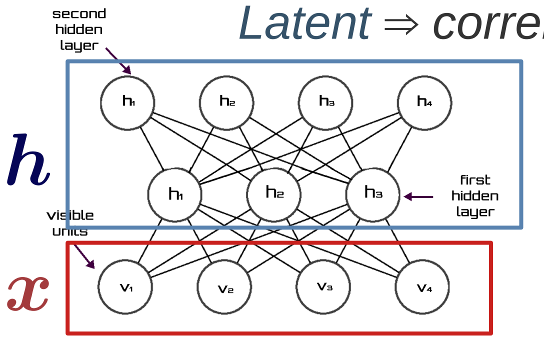
$$\boxed{\mathcal{L}(\boldsymbol{\theta}|X) = \sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x} = x^{(m)}\right)}$$

Gradient ascent $\nabla_{\boldsymbol{\theta}} \mathcal{L}$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \left. \frac{\partial \mathcal{L}}{\partial \theta_i} \right|_{\theta = \theta_i^{(t)}}$$

Training Energy-Based Models (EBMs)

Latent ⇒ *correlations*



Visible ⇒ *data*

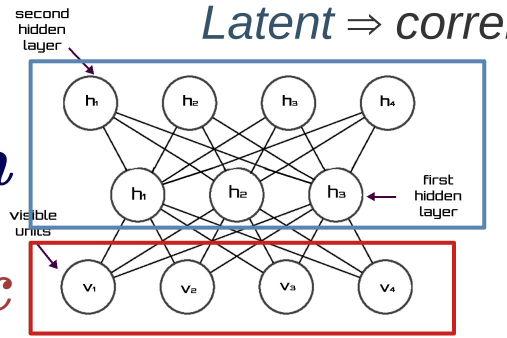
Marginal distribution

$$\mathcal{E}(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta}) \Rightarrow p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{h})}}{Z_{\boldsymbol{\theta}}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Training Energy-Based Models (EBMs)

Latent ⇒ *correlations*

$\boldsymbol{h}$



$\boldsymbol{x}$

Visible ⇒ *data*

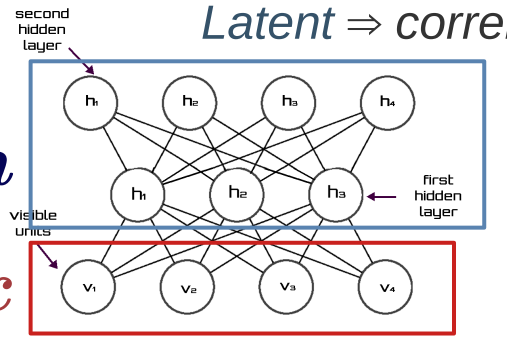
Marginal distribution

$$\mathcal{E}(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta}) \Rightarrow p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{h})}}{Z_{\boldsymbol{\theta}}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Training: maximize the log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}|X) = \langle \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \rangle_{p_{\mathrm{data}}} = \langle -E_{\boldsymbol{\theta}}(\boldsymbol{x}) \rangle_{p_{\mathrm{data}}} - \log Z_{\boldsymbol{\theta}}$$

Training Energy-Based Models (EBMs)



Latent ⇒ *correlations*

$\boldsymbol{h}$

$\boldsymbol{x}$

Visible ⇒ *data*

Marginal distribution

$$\mathcal{E}(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta}) \Rightarrow p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{h})}}{Z_{\boldsymbol{\theta}}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Training: maximize the log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}|X) = \langle \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \rangle_{p_{\mathrm{data}}} = \langle -E_{\boldsymbol{\theta}}(\boldsymbol{x}) \rangle_{p_{\mathrm{data}}} - \log Z_{\boldsymbol{\theta}}$$

$$\boldsymbol{\nabla}\mathcal{L} = \boxed{\langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\mathrm{data}}}} - \boxed{\langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}}$$

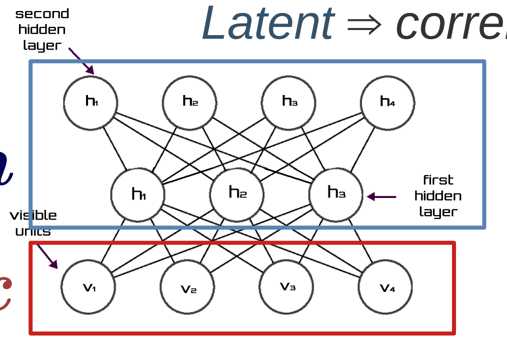
Easy

Hard ⇒ MCMC sampling

(Stochastic) gradient ascent

Training Energy-Based Models (EBMs)

Latent ⇒ *correlatic*



Visible ⇒ *data*

T

$$\mathcal{L}(\boldsymbol{\theta}|X) =$$

Insufficient Monte Carlo samplings have strong effects on the quality of the model learned

- Decelle, Furtlehner, Seoane NeurIPS (2021)
- Agoritsas, Catania, Decelle, Seoane ICML (2023)
- Carbone, Decelle, Seoane, Rosset, arXiv: 2307.06797 (2023)
- Béreux, Decelle, Furtlehner, Seoane - SciPost Physics (2023)

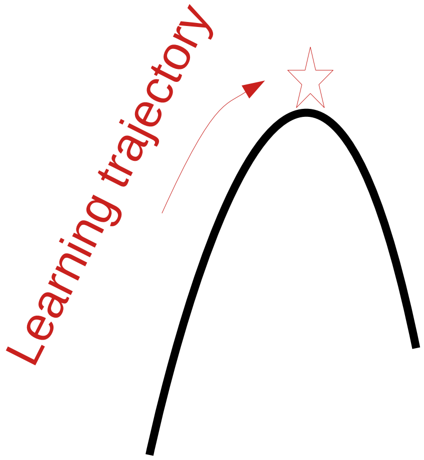
$$\boldsymbol{\nabla}\mathcal{L} = \boxed{\langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\mathrm{data}}}} - \boxed{\langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}}$$

(Stochastic) gradient ascent

Easy

Hard ⇒ MCMC sampling

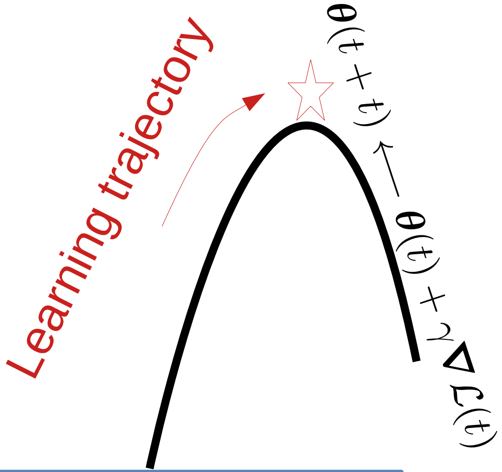
On the gradient ascent



Update rule: $\nabla \mathcal{L}_{\boldsymbol{\theta}} = \langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\mathrm{data}}} - \langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\Theta}}}$

$$\boldsymbol{\theta}(t + t) \longleftarrow \boldsymbol{\theta}(t) + \gamma \nabla \mathcal{L}(t)$$

On the gradient ascent

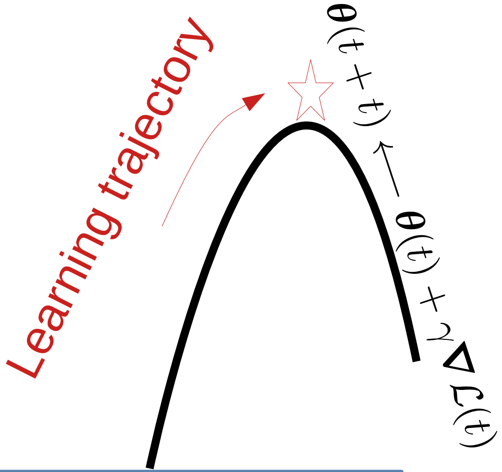


Update rule: $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\mathrm{data}}} - \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\Theta}}}$

Fixed point : $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0} \quad \Rightarrow \quad \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathrm{data}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \quad \forall \theta_i$

Moment matching statistics

On the gradient ascent



Update rule: $\boldsymbol{\nabla}\mathcal{L}_{\boldsymbol{\theta}} = \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\mathrm{data}}} - \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\Theta}}}$

Fixed point : $\boldsymbol{\nabla}\mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0} \Rightarrow \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathrm{data}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \quad \forall \theta_i$

Moment matching statistics

If the optimization problem is convex, as e.g. $\boldsymbol{\nabla} E_{\boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{x})$

$$E = -\sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i \Rightarrow \begin{cases} \langle S_i S_j \rangle_{p_{\boldsymbol{J},h}} = \langle S_i S_j \rangle_{p_{\mathrm{data}}} & \forall i,j \\ \langle S_i \rangle_{p_{\boldsymbol{J},h}} = \langle S_i \rangle_{p_{\mathrm{data}}} \end{cases}$$

Generating new samples

Empirical *Model*

$$p_{\text{data}}(x) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Dominated minimum free-energy configurations

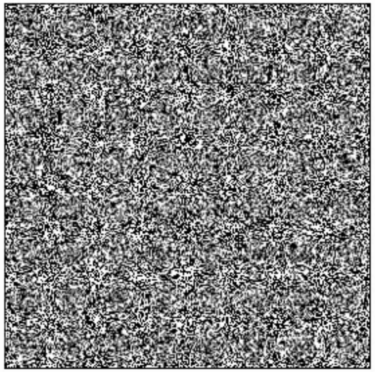
$$\{\boldsymbol{x}\}_{\text{eq},\boldsymbol{\theta}} \sim \mathcal{D}$$

$\Longrightarrow$

Markov Chain Monte Carlo
Langevin dynamics

Generate new samples

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\text{data}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \quad \forall \theta_i$$



Generating new samples

Empirical *Model*

$$p_{\mathrm{data}}(x) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Dominated minimum free-energy configurations

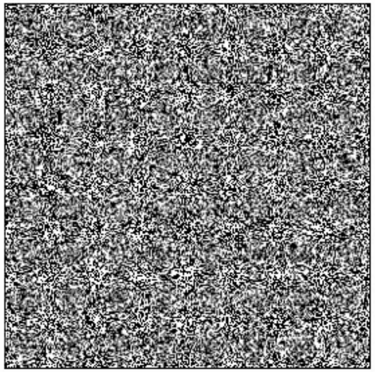
$$\{\boldsymbol{x}\}_{\mathrm{eq},\boldsymbol{\theta}} \sim \mathcal{D}$$

$\Rightarrow$

Markov Chain Monte Carlo
Langevin dynamics

Generate new samples

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathrm{data}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \quad \forall \theta_i$$



$E_{\boldsymbol{\theta}}(\boldsymbol{x})$ *Effective* model for the data

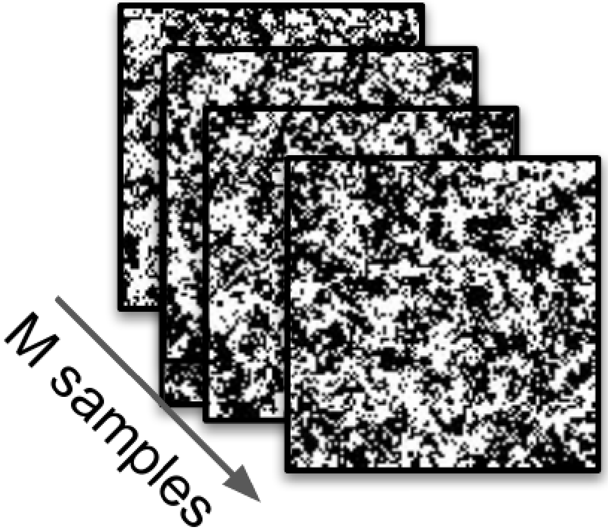
$\Rightarrow$ *Free-energy landscape*

Modeling, <u>interpretability</u>

Interpreting the energy function

Inverse Ising problem

Nguyen, H. C., Zecchina, R., & Berg, J. (2017) Advances in Physics



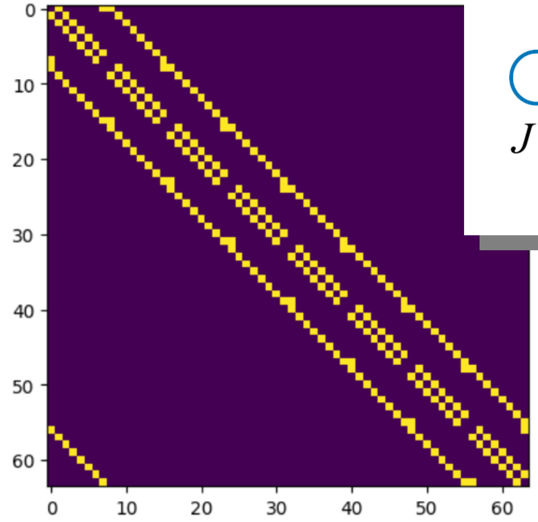
Dataset

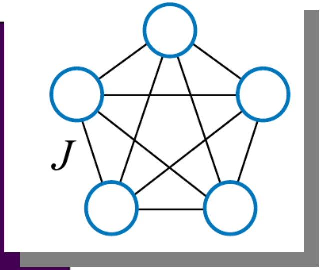
$$E_{\text{Ising 2D}}(\boldsymbol{S}) = -\hat{J}\sum_{\langle i,j\rangle} S_i S_j$$

$$\hat{\beta} = 1/\hat{T}$$

Am I able to infer which was the interaction model that generated it?

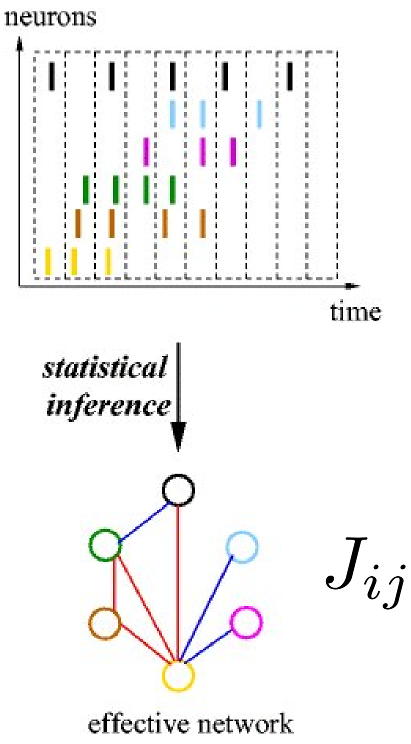
$$E_{J,h}(\boldsymbol{x}) = -\sum_{ij} J_{ij}x_i x_j - \sum_i h_i x_i$$

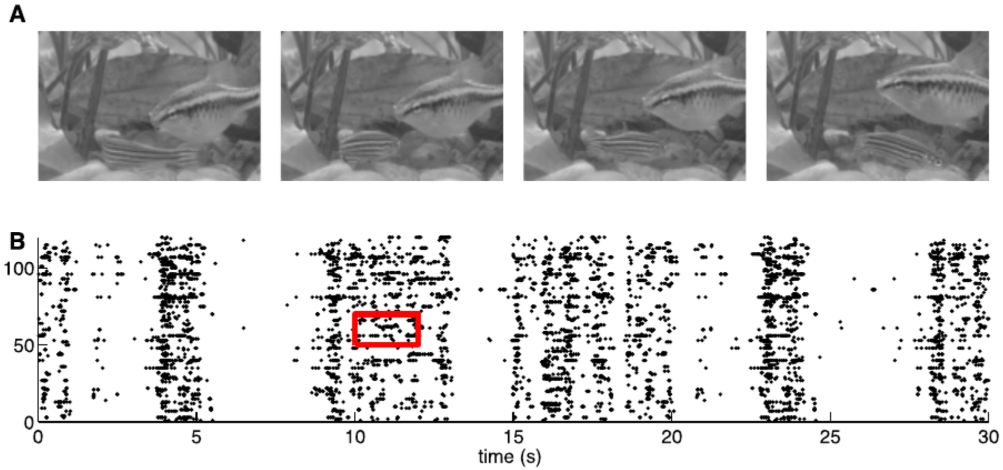


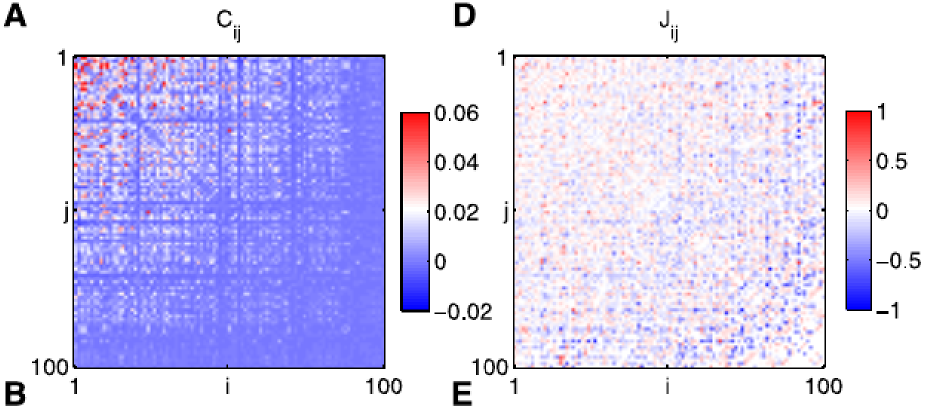


Applications I: reconstruction of neural connections

Tavoni, G., Cocco, S., & Monasson, R. (2016)





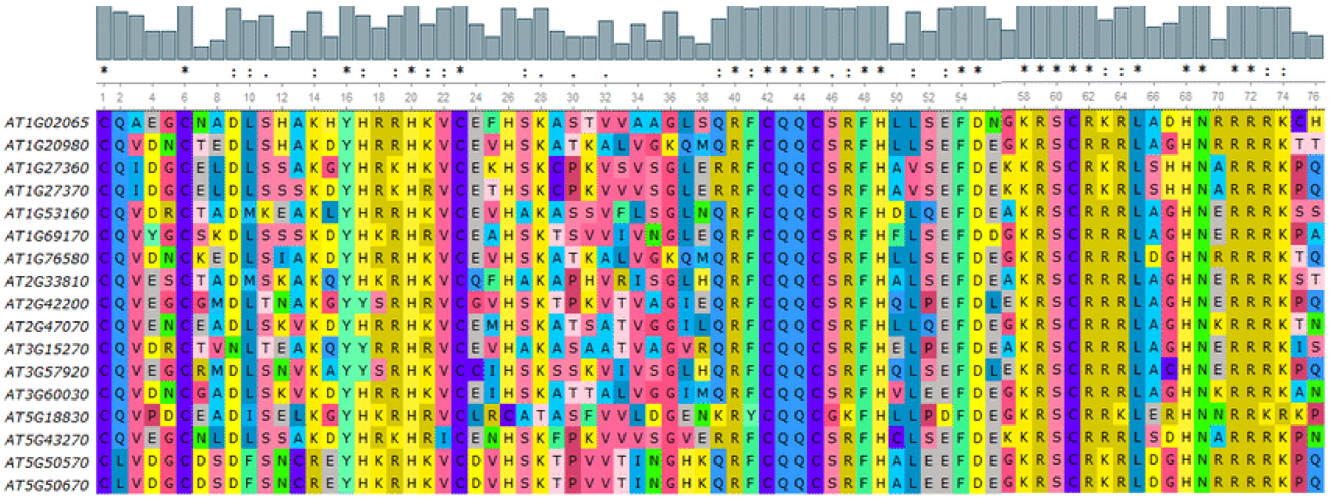


Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., & Berry, M. J. (2014).

Roudi, Y., Aurell, E., & Hertz, J. A. (2009)
Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006)

Applications II: Inverse Potts
Direct coupling analysis (DCA)

$$E_{J,h}(\boldsymbol{S}) = -\sum_{i,j=1}^{N_v} \sum_{q_1,_2=1}^{N_q} J_{ij}^{q_1,q_2} \delta_{S_i,q_1} \delta_{S_i,q_2} - \sum_{i=1}^{N_v} \sum_{q=1}^{N_q} h_i^q \delta_{S_i,q} \qquad S_i = \{1, \ldots, q\}$$



MSA

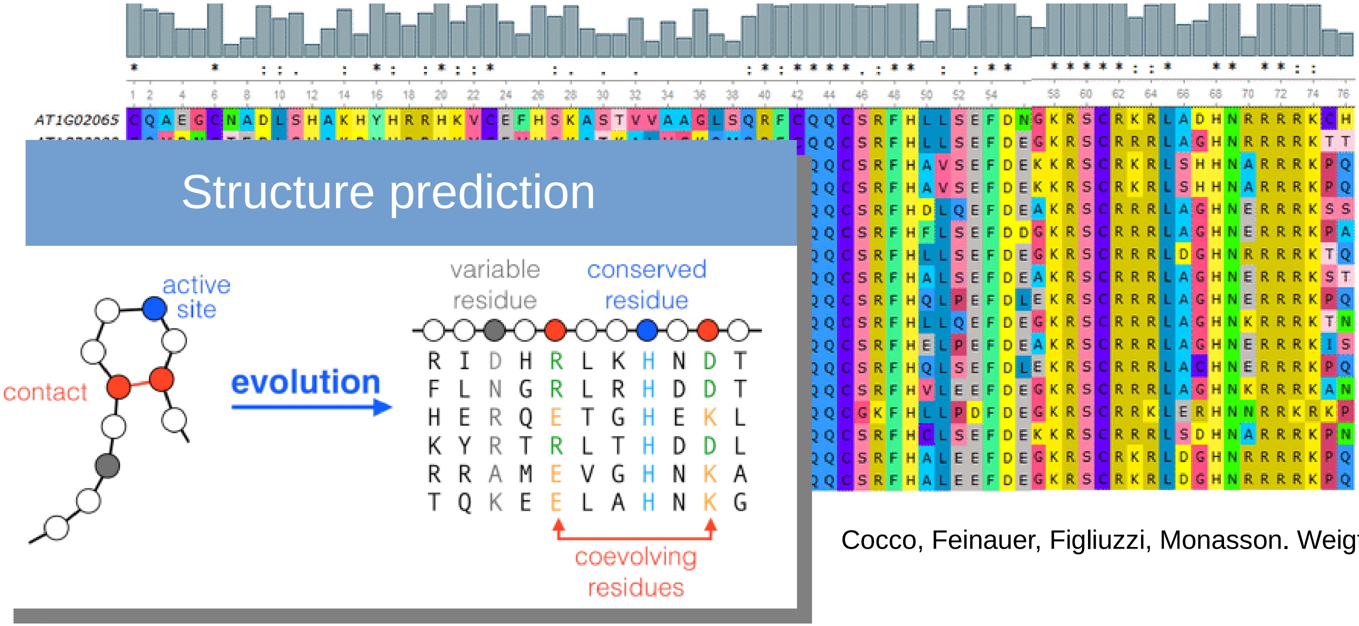
q=21

Model the “true” *fitness landscape*

Statistical sequence landscape

Applications II: Inverse Potts
Direct coupling analysis (DCA)

$$E_{J,h}(\boldsymbol{S}) = -\sum_{i,j=1}^{N_v} \sum_{q_1,_2=1}^{N_q} J_{ij}^{q_1,q_2} \delta_{S_i,q_1} \delta_{S_i,q_2} - \sum_{i=1}^{N_v} \sum_{q=1}^{N_q} h_i^q \delta_{S_i,q} \qquad S_i = \{1,\ldots,q\}$$



MSA

q=21

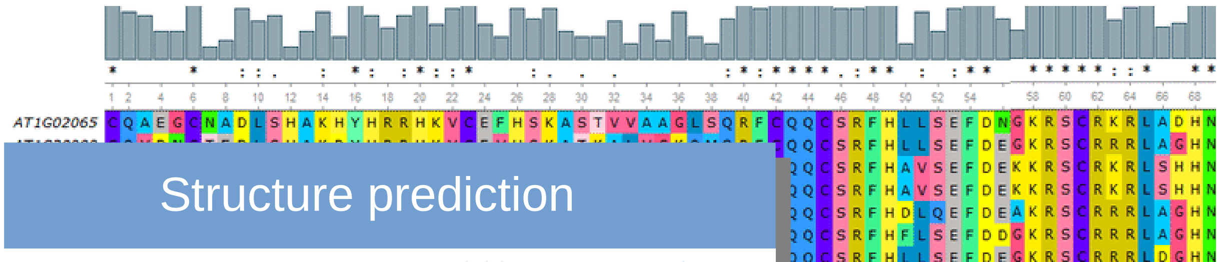
Model the “true”
fitness landscape

Statistical sequence landscape

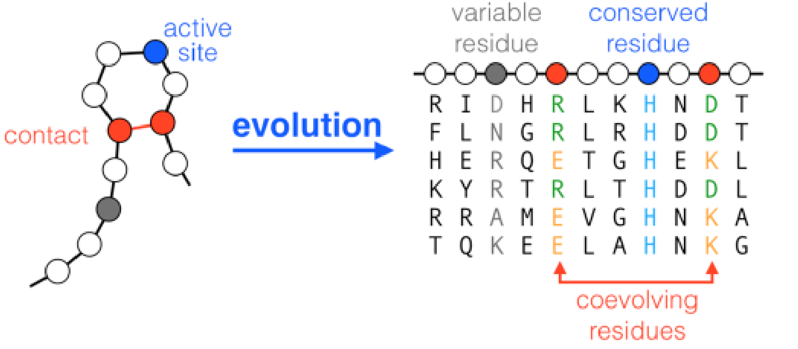
Cocco, Feinauer, Figliuzzi, Monasson. Weigt, Rep. Prog. Phys. 81 (2018) 032601

Ex. Inverse Potts
Direct coupling analysis (DCA)

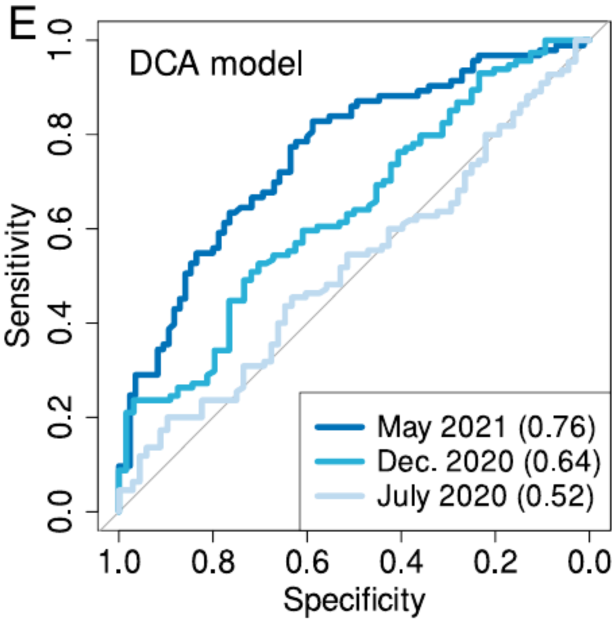
$$E_{J,h}(S) = -\sum_{i,j=1}^{N_v}\sum_{q_1,_2=1}^{N_q} J_{ij}^{q_1,q_2}\,\delta_{S_i,q_1}\delta_{S_i,q_2} - \sum_{i=1}^{N_v}\sum \cdots q\}$$



Structure prediction



Mutation prediction



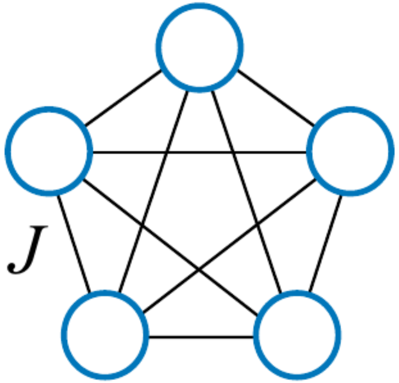
Cocco, Feinauer, Figliuzzi, Monasson. Weigt, Rep. Prog. Phys. 81 (2018) 032601

Rodriguez-Rivas, J., Croce, G., Muscat, M., & Weigt, M.
Proceedings of the National Academy of Sciences, (2022).

Pairwise models : The Boltzmann machine

Hinton and Sejnowski (1983)

$$E_{J,h}(\boldsymbol{x}) = -\sum_{ij} J_{ij} x_i x_j - \sum_i h_i x_i$$



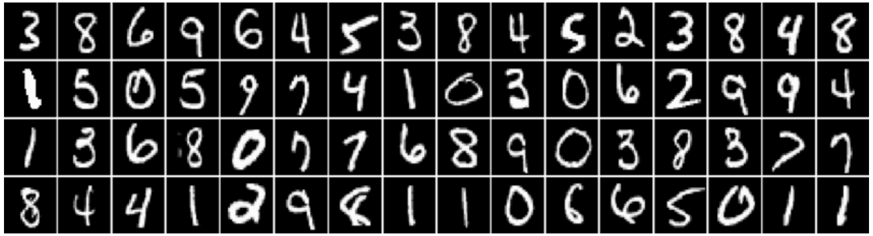
Simple and easy to interpret, but are strongly limited...

Pairwise models : The Boltzmann machine

Hinton and Sejnowski (1983)

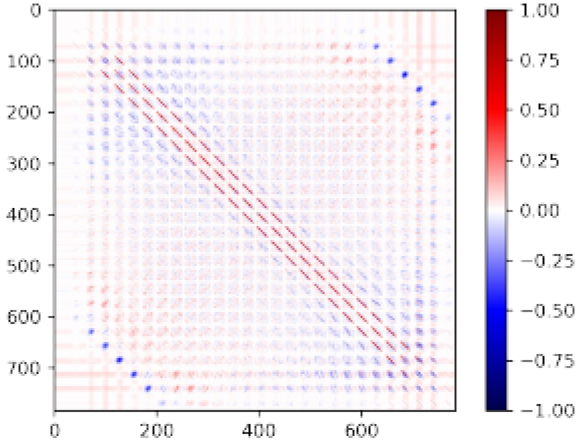
$$E_{J,h}(\boldsymbol{x}) = -\sum_{ij} J_{ij} x_i x_j - \sum_{i} h_i x_i$$

Simple and easy to interpret, but are strongly limited...



learning

BM inferred pairwise coupling matrix



Pairwise models : The Bolt

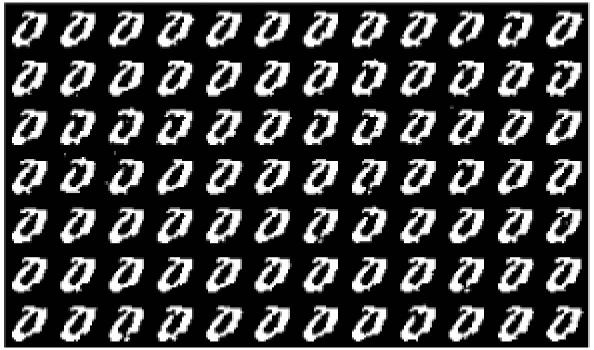
We need to encode **higher order correlations** !

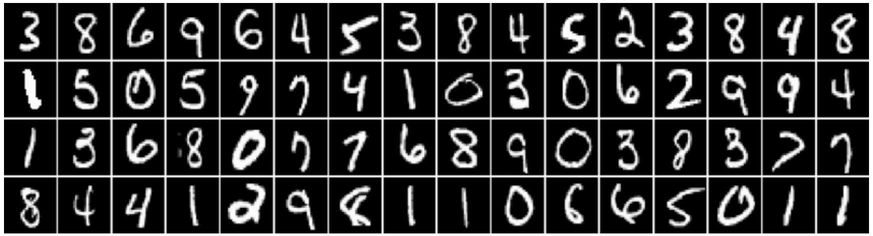
$$E_{J,h}(\boldsymbol{x}) = -\sum_{ij} J_{ij} x_i x_j - \sum_{i} h_i x_i$$

Simple and easy to interpret, but are strongly limited...

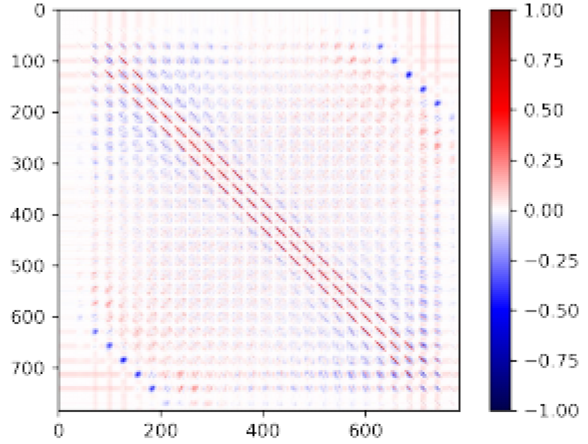
Generation

Samples generated with the BM

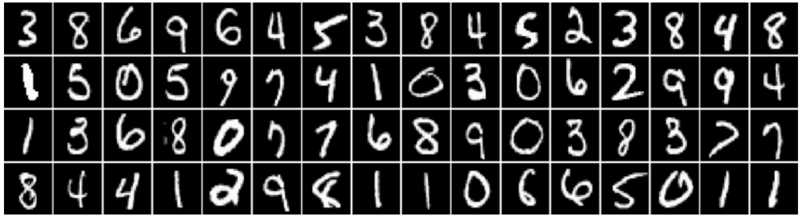


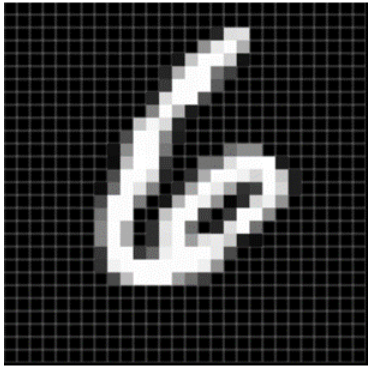


learning



Encoding high-order correlations





$$f_i = \langle x_i \rangle_{\text{data}}$$

$$f_{ij} = \langle x_i x_j \rangle_{\text{data}}$$

$$f_{ijk} = \langle x_i x_j x_k \rangle_{\text{data}}$$

$$f_{i_1 \cdots i_n} = \langle x_{i_1} \cdots x_{i_n} \rangle_{\text{data}}$$

parameters diverge too fast...

$$E(\boldsymbol{x}) = -\sum_i h_i x_i - \sum_{ij} J^{(2)}_{ij} x_i x_j - \sum_{ijk} J^{(3)}_{ijk} x_i x_j x_k - \sum_{ijkl} J^{(4)}_{ijkl} x_i x_j x_k x_l + \cdots$$

Encoding high-order correlations

But in real data the interactions are sparse

Only some *n-tuples* of variables are correlated

$$f_i = \langle x_i \rangle_{\text{data}}$$

$$f_{ij} = \langle x_i x_j \rangle_{\text{data}}$$

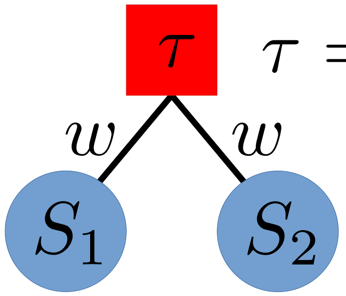
$$f_{ijk} = \langle x_i x_j x_k \rangle_{\text{data}}$$

$$f_{i_1 \cdots i_n} = \langle x_{i_1} \cdots x_{i_n} \rangle_{\text{data}}$$

parameters diverge too fast...

$$E(\boldsymbol{x}) = -\sum_i h_i x_i - \sum_{ij} J^{(2)}_{ij} x_i x_j - \sum_{ijk} J^{(3)}_{ijk} x_i x_j x_k - \sum_{ijkl} J^{(4)}_{ijkl} x_i x_j x_k x_l + \cdots$$

Alternative solution: add hidden variables



$\tau = \pm 1$

$S_i = \pm 1$

$$\mathcal{H}(S_1, S_2, \tau) = -w\tau(S_1 + S_2)$$

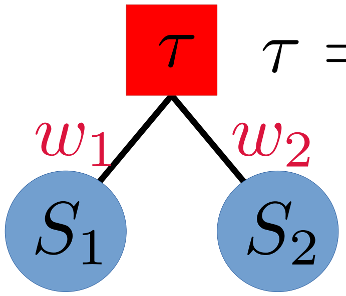
Marginal probability $p(S_1, S_2) = \dfrac{e^{-\mathcal{H}(S_1,S_2)}}{Z}$

$$\mathcal{H} = -\log \sum_{\tau=\pm 1} e^{w\tau(S_1+S_2)} = -\log 2\cosh\left[w(S_1 + S_2)\right]$$

$$= -JS_1S_2 - J$$

$$\Rightarrow \boxed{\cosh 2w = e^{2J}} \qquad J > 0$$

Alternative solution: add hidden variables



$\tau = \pm 1$

$S_i = \pm 1$

$$\mathcal{H}(S_1, S_2, \tau) = -\tau(w_1 S_1 + w_2 S_2)$$

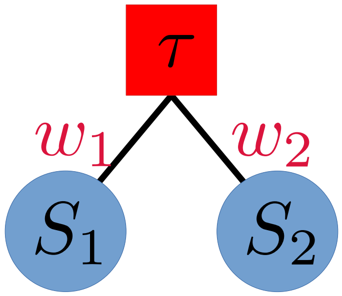
Marginal probability $$p(S_1, S_2) = \frac{e^{-\mathcal{H}(S_1,S_2)}}{Z}$$

$$\mathcal{H} = -\log \sum_{\tau=\pm 1} e^{\tau(w_1 S_1 + w_2 S_2)} = -\log 2\cosh\left[w_1 S_1 + w_2 S_2\right]$$
$$= -J S_1 S_2 - J$$

The encoding is not unique !

$$\Rightarrow \boxed{\frac{\cosh(w_1 + w_2)}{\cosh(w_1 - w_2)} = e^{2J}} \quad J > 0$$

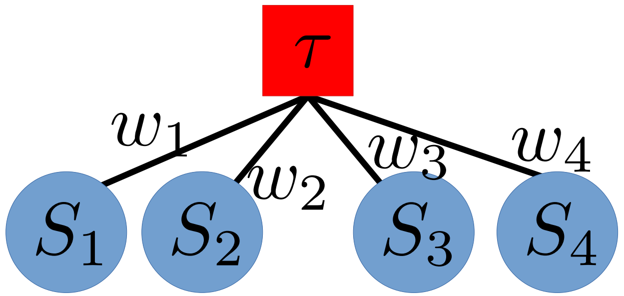
Alternative solution: add hidden variables



$$\mathcal{H}(S_1, S_2, \tau) = -\tau(w_1 S_1 + w_2 S_2 + \theta) + h_1 S_1 + h_2 S_2$$

There are even more ways to encode the same interaction if you consider biases…

Alternative solution: add hidden variables



$$\mathcal{H}(S_1, S_2, \tau) = -\tau(w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4)$$

$$\mathcal{H}(S_1, S_2, S_3, S_4) = -\log 2 \cosh \left[w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4\right]$$

$$= -J^{(4)}_{1234} S_1 S_2 S_3 S_4 - J^{(2)}_{12} S_1 S_2 - J^{(2)}_{13} S_1 S_3 - J^{(2)}_{14} S_1 S_4 - J^{(2)}_{23} S_2 S_3 - J^{(2)}_{24} S_2 S_4 - J^{(2)}_{34} S_3 S_4 + C$$

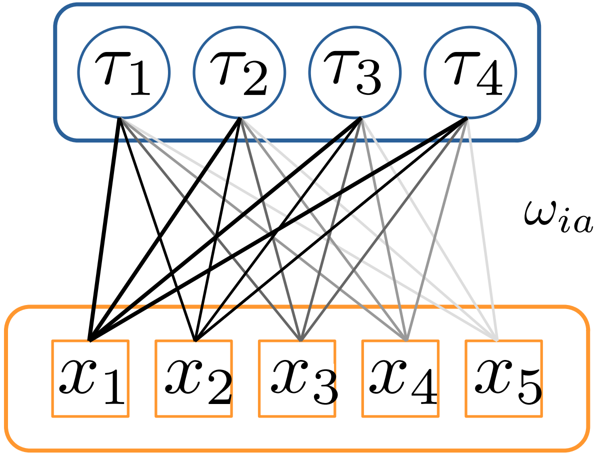
In order to encode an interaction model with at most *k*-body interactions we need O(N_k) hidden nodes, with N_k the number of non-zero $J^{(k)}$ couplings (# parameters O(N_k)N) << O(N^k))

The Restricted Boltzmann Machine

-Smolensky, P. (1986)

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\tau}) = -\sum_{ia} x_i w_{ia} \tau_a - \sum_i \eta_i x_i - \sum_a \theta_a \tau_a$$

$$\boldsymbol{\theta} = \{W, \eta, \theta\}$$



Visible : **data**

Hidden : “Neurons” → **features extracted**

Universal approximator !

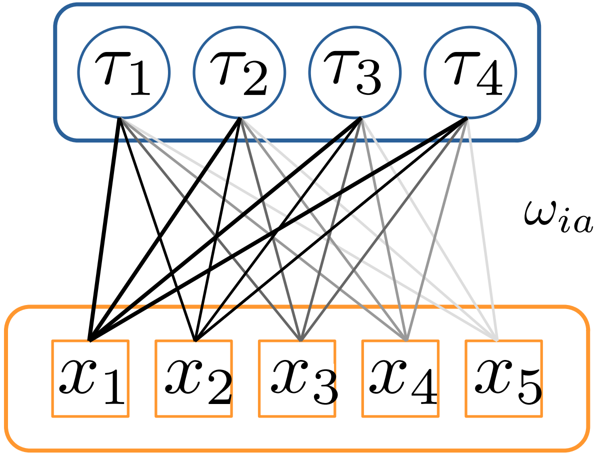
Le Roux and Bengio. Neural computation (2008)

The Restricted Boltzmann Machine

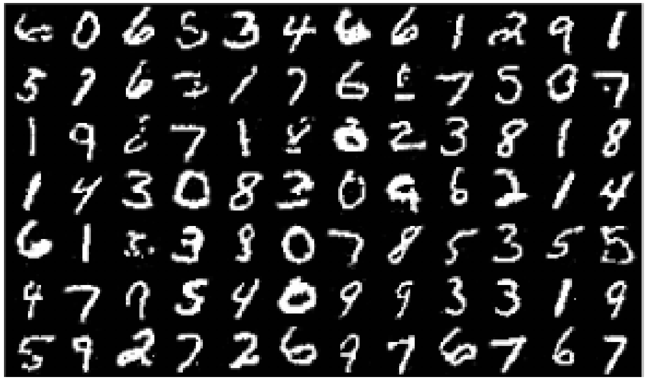
-Smolensky, P. (1986)

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\tau}) = -\sum_{ia} x_i w_{ia} \tau_a - \sum_i \eta_i x_i - \sum_a \theta_a \tau_a$$

$$\boldsymbol{\theta} = \{W, \eta, \theta\}$$



B3 **Samples generated with the RBM**



Universal approximator !

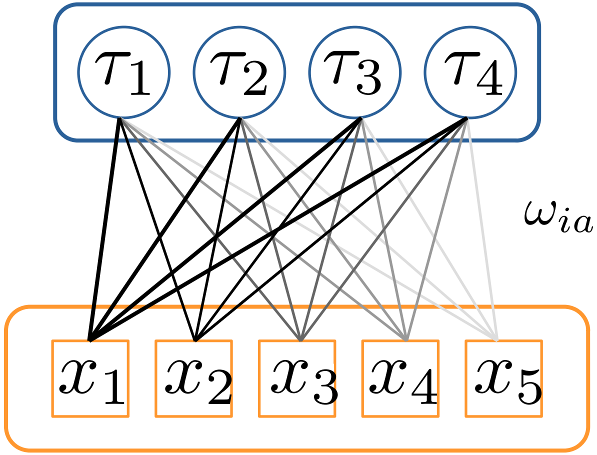
Le Roux and Bengio. Neural computation (2008)

The Restricted Boltzmann Machine

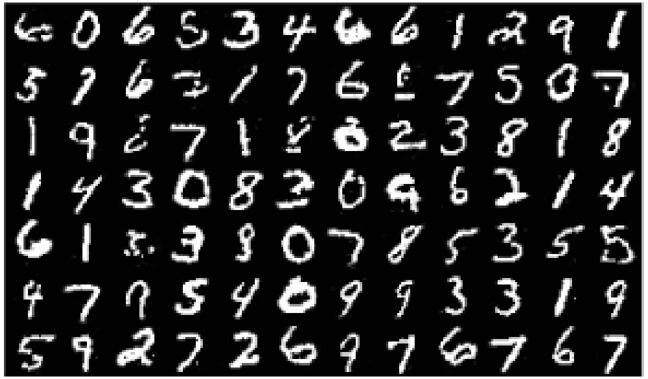
-Smolensky, P. (1986)

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\tau}) = -\sum_{ia} x_i w_{ia} \tau_a - \sum_{i} \eta_i x_i - \sum_{a} \theta_a \tau_a$$

$$\boldsymbol{\theta} = \{W, \eta, \theta\}$$



B3 Samples generated with the RBM



The RBM is **much more expressive** than the BM, but can we **make it just as interpretable**?

$$
\begin{aligned}
\mathcal{H}_{RBM}(\boldsymbol{v}) &= -\sum_j b_j v_j - \sum_i \ln\left(1 + e^{c_i + \sum_j W_{ij} v_j}\right) \\
&= -\sum_j H_j v_j - \sum_{j_1 > j_2} J^{(2)}_{j_1 j_2} v_{j_1} v_{j_2} - \sum_{j_1 > j_2 > j_3} J^{(3)}_{j_1 j_2 j_3} v_{j_1} v_{j_2} v_{j_3} + \dots
\end{aligned}
$$

The RBM as a model for interacting spins

From the RBM to a generalized Ising model

$$E_{\boldsymbol{\theta}}^{\mathrm{RBM}}(\boldsymbol{v}) = -\log\left(\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}\right)$$

The RBM

Rewrite in terms of

$\boldsymbol{\sigma}, \boldsymbol{\tau} \quad \sigma_j, \tau_i \in \{\pm 1\}$

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_j \eta_j \sigma_j - \sum_i \ln\cosh\left(\sum_j w_{ij}\sigma_j + \theta_i\right).$$

From the RBM to a generalized Ising model

$$E^{\mathrm{RBM}}_{\boldsymbol{\theta}}(\boldsymbol{v}) = -\log\left(\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}\right)$$

The RBM

Rewrite in terms of

$\boldsymbol{\sigma}, \boldsymbol{\tau} \quad \sigma_j, \tau_i \in \{\pm 1\}$

$$\begin{aligned}
\mathcal{H}(\boldsymbol{\sigma}) &= -\sum_j \eta_j \sigma_j - \sum_i \ln\cosh\left(\sum_j w_{ij}\sigma_j + \zeta_i\right).\\
&= -\sum_j \eta_j \sigma_j - \sum_{\boldsymbol{\sigma}'} \prod_j \delta_{\sigma_j \sigma'_j} \sum_i \ln\cosh\left(\sum_j w_{ij}\sigma'_j + \zeta_i\right).\\
&= -\sum_j \eta_j \sigma_j - \frac{1}{2^{N_{\mathrm{v}}}} \sum_{\boldsymbol{\sigma}'} \prod_j \left(1 + \sigma_j \sigma'_j\right) \sum_i \ln\cosh\left(\sum_j w_{ij}\sigma'_j + \zeta_i\right).\\
&= -\sum_j H_j \sigma_j - \sum_{j_1 > j_2} J^{(2)}_{j_1 j_2} \sigma_{j_1}\sigma_{j_2} - \sum_{j_1 > j_2 > j_3} J^{(3)}_{j_1 j_2 j_3} \sigma_{j_1}\sigma_{j_2}\sigma_{j_3} - \cdots
\end{aligned}$$

From the RBM to a generalized Ising model

$$E_{\boldsymbol{\theta}}^{\mathrm{RBM}}(\boldsymbol{v}) = -\log\left(\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}\right)$$

The RBM

Rewrite in terms of

$$\boldsymbol{\sigma}, \boldsymbol{\tau} \qquad \sigma_j, \tau_i \in \{\pm 1\}$$

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_j H_j \sigma_j - \sum_{j_1>j_2} J^{(2)}_{j_1 j_2} \sigma_{j_1}\sigma_{j_2} - \sum_{j_1>j_2>j_3} J^{(3)}_{j_1 j_2 j_3} \sigma_{j_1}\sigma_{j_2}\sigma_{j_3} - \cdots$$

Given an RBM, we know which effective Ising Model it corresponds to

$$\boxed{\begin{aligned} H_j &= \eta_j + \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \sum_i \sigma'_j \ln\cosh\left(\sum_k w_{ik}\sigma'_k + \zeta_i\right) \\ J^{(n)}_{j_1\dots j_n} &= \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \sum_i \sigma'_{j_1}\dots\sigma'_{j_n} \ln\cosh\left(\sum_k w_{ik}\sigma'_k + \zeta_i\right) \end{aligned}}$$

From the RBM to a generalized Ising model

Introduce the random variable

$$X_i^{(j_1 \dots j_n)} \equiv \sum_{\mu=n+1}^{N_\mathrm{v}} w_{ij_\mu} \sigma'_{j_\mu}$$

Central limit theorem

$$H_j = \eta_j + \frac{1}{2} \sum_i \mathbb{E}_{X_i^{(j)}} \left[\ln \frac{\cosh\left(\zeta_i + w_{ij} + X_i^{(j)}\right)}{\cosh\left(\zeta_i - w_{ij} + X_i^{(j)}\right)} \right]$$

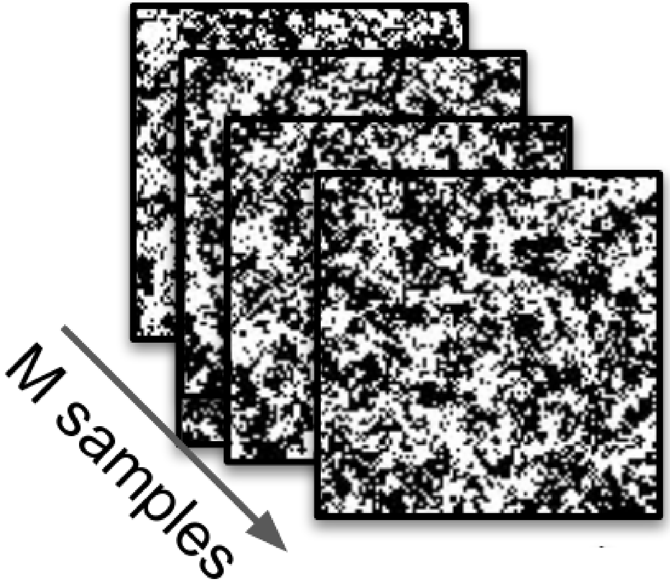
$$J^{(2)}_{j_1 j_2} = \frac{1}{4} \sum_i \mathbb{E}_{X_i^{(j_1 j_2)}} \left[\ln \frac{\cosh\left(\zeta_i + w_{ij_1} + w_{ij_2} + X_i^{(j_1 j_2)}\right) \times \cosh\left(\zeta_i - (w_{ij_1} + w_{ij_2}) + X_i^{(j_1 j_2)}\right)}{\cosh\left(\zeta_i + (w_{ij_1} - w_{ij_2}) + X_i^{(j_1 j_2)}\right) \times \cosh\left(\zeta_i - (w_{ij_1} - w_{ij_2}) + X_i^{(j_1 j_2)}\right)} \right]$$

Numerical controlled experiments

$$H_{\mathrm{original}}(\boldsymbol{\sigma}) = -\sum_i h_i^* \sigma_i - \sum_{ij} J_{ij}^{*(2)} \sigma_i \sigma_j - \left(-\sum_{ijk} J_{ijk}^{*(3)} \sigma_i \sigma_j \sigma_k \right)$$

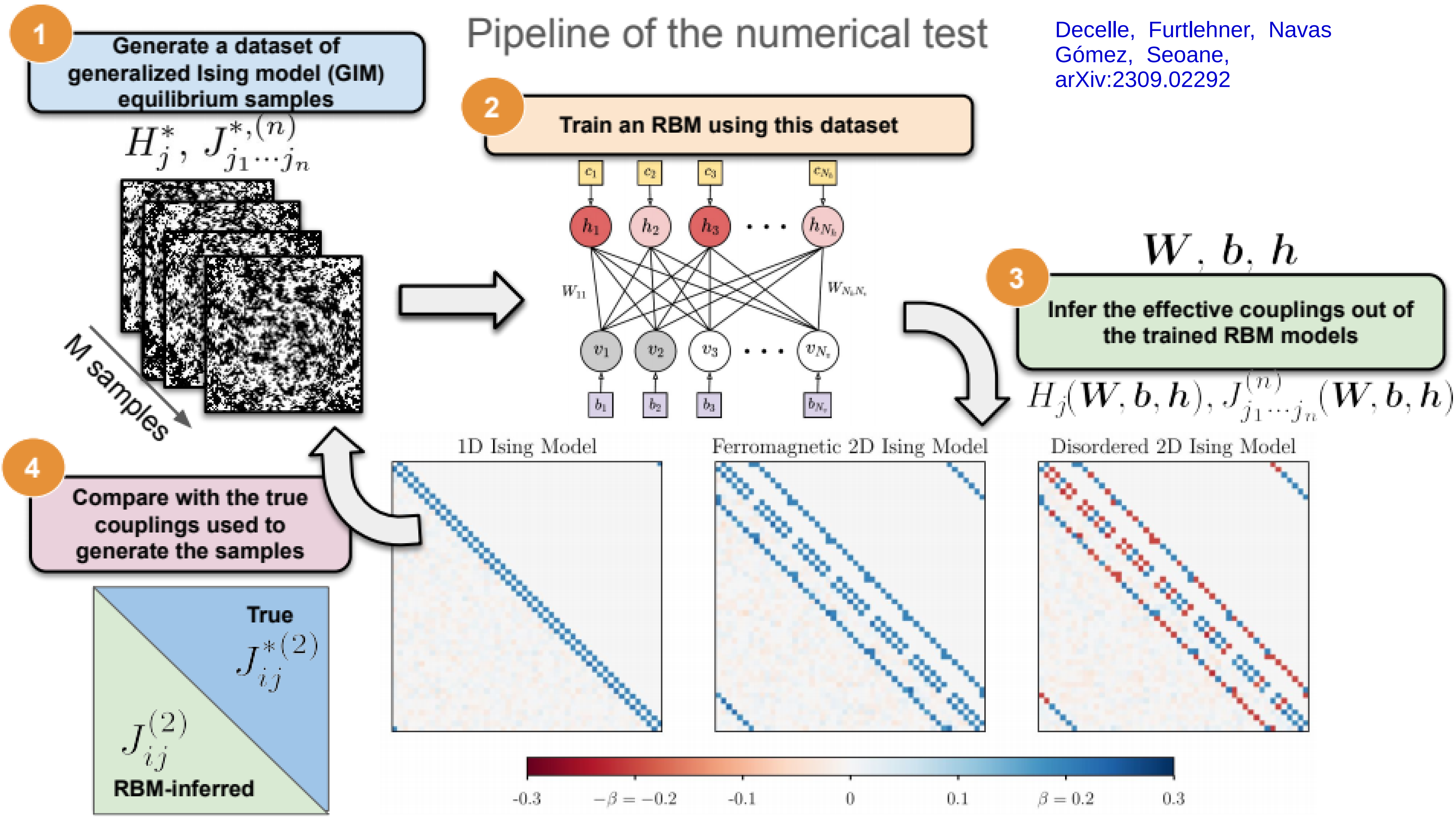
$$\beta = \frac{1}{T}$$

Generate equilibrium samples
With a known model

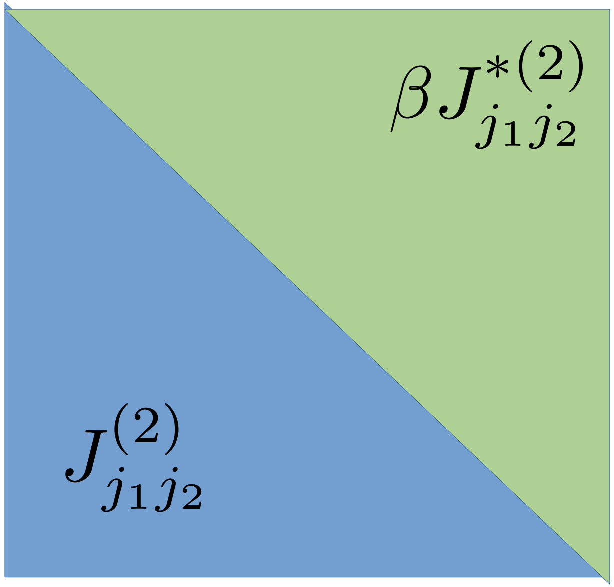


Pipeline of the numerical test

Decelle, Furtlehner, Navas Gómez, Seoane, arXiv:2309.02292



“Experimental test”



Coupling matrix used to generate the samples

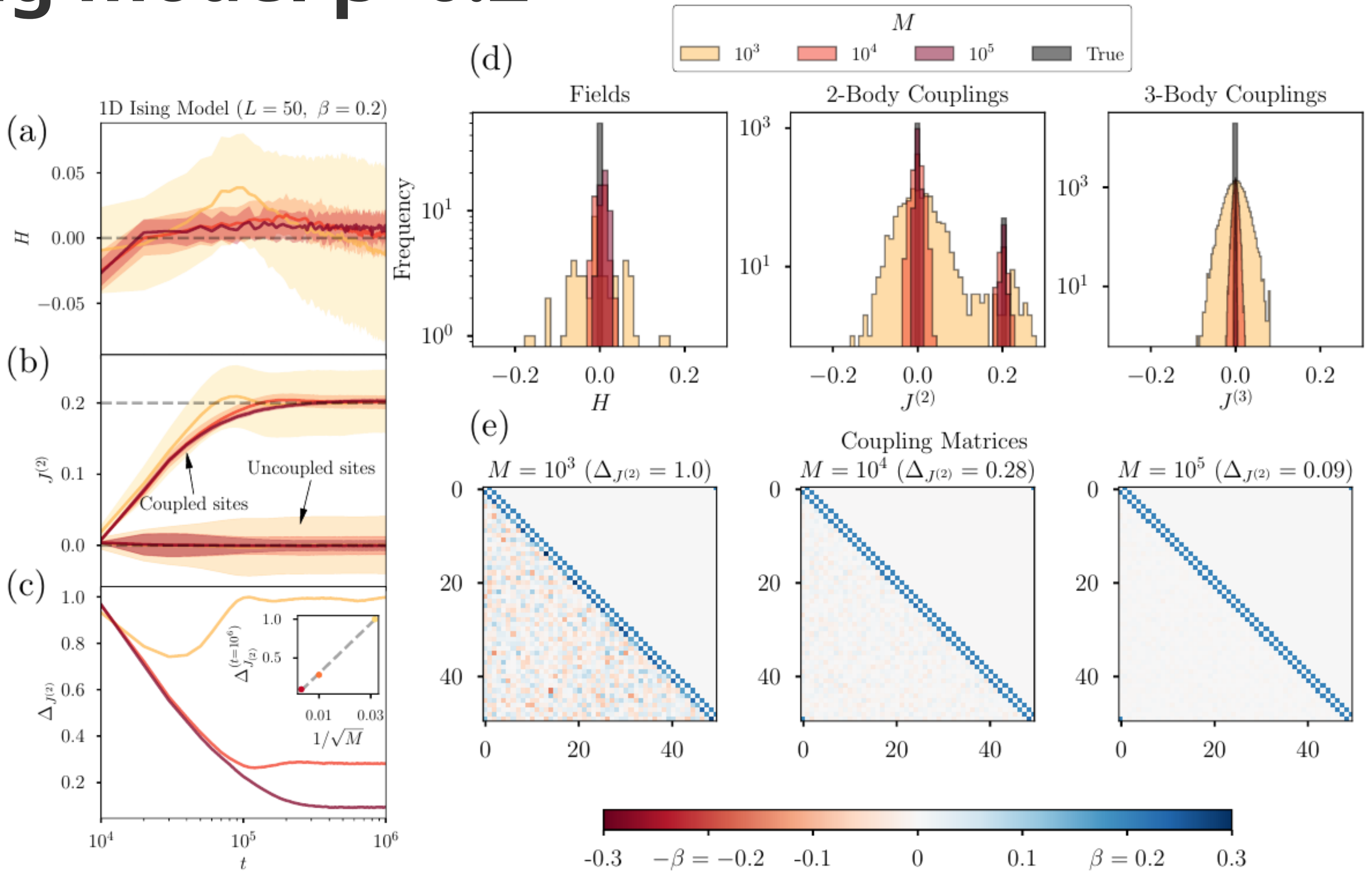
Inferred coupling matrix

$$\Delta_{J^{(2)}} = \sqrt{\frac{\sum_{j_1>j_2}\left(J^{(2)}_{j_1 j_2} - \beta J^{*(2)}_{j_1 j_2}\right)^2}{\sum_{j_1>j_2}\beta J^{*(2)}_{j_1 j_2}{}^2}}.$$

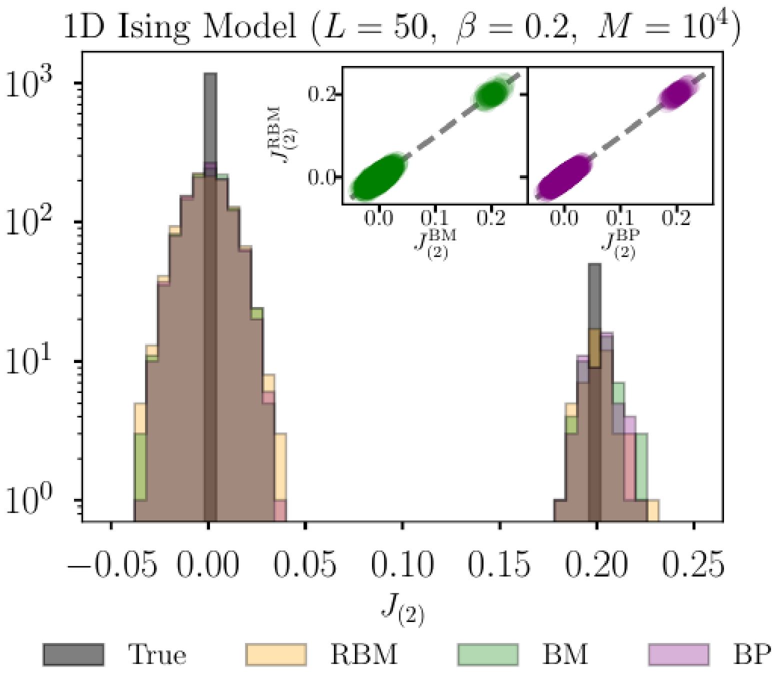
We want to recover:

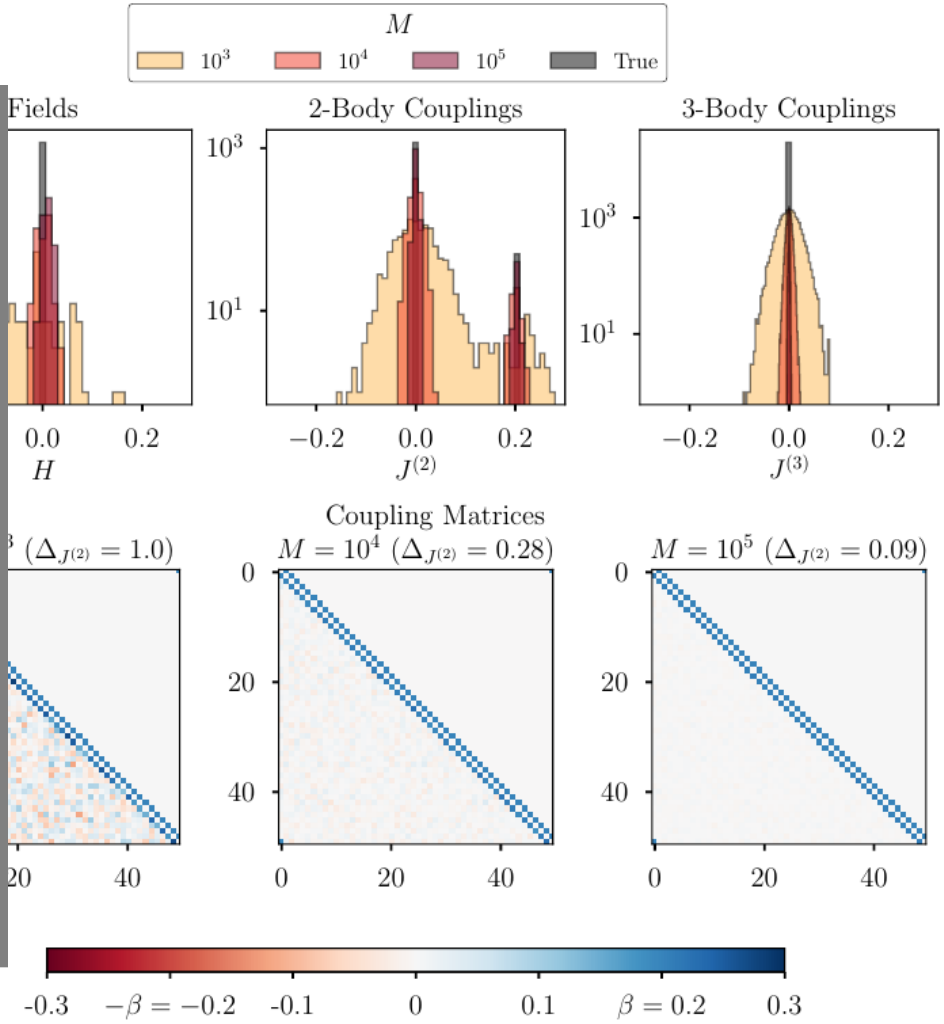
- The connectivity network
- The coupling strength

1D Ising model β=0.2



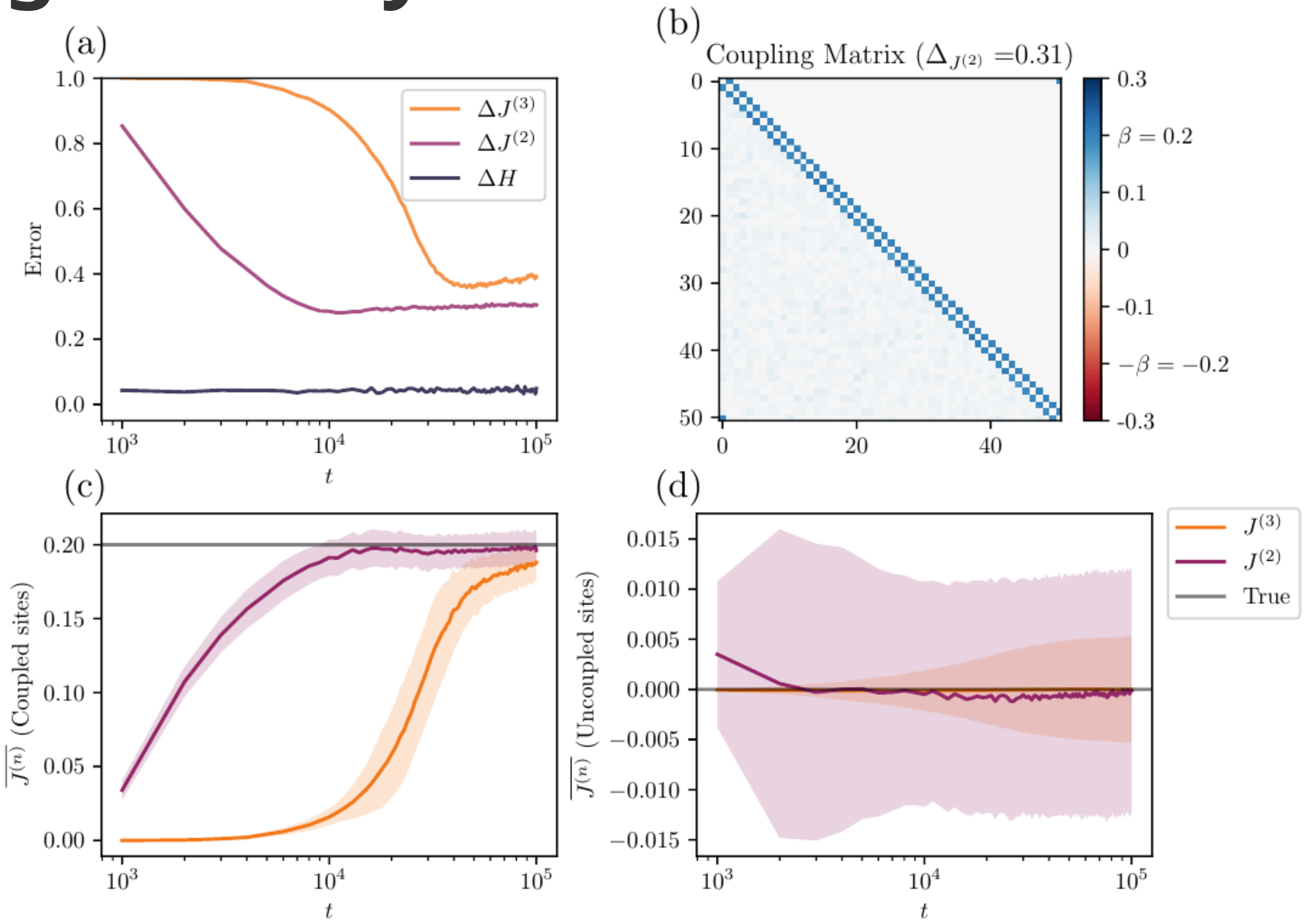
1D Ising model β=0.2



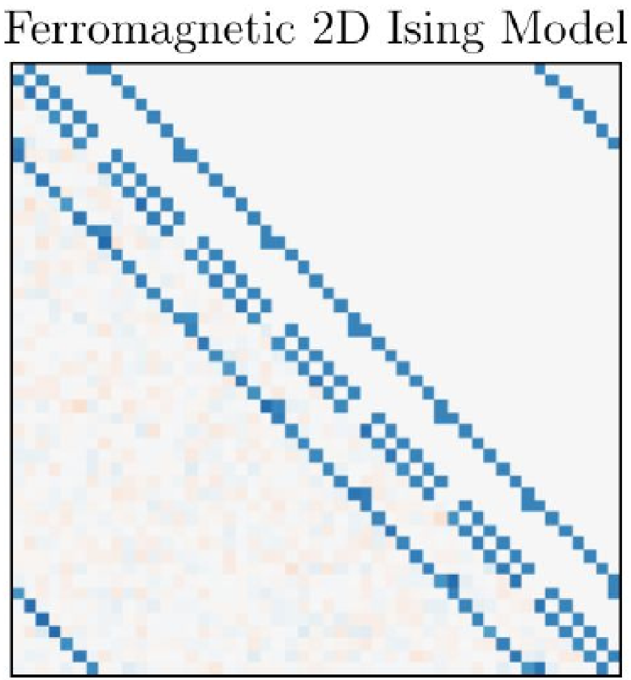


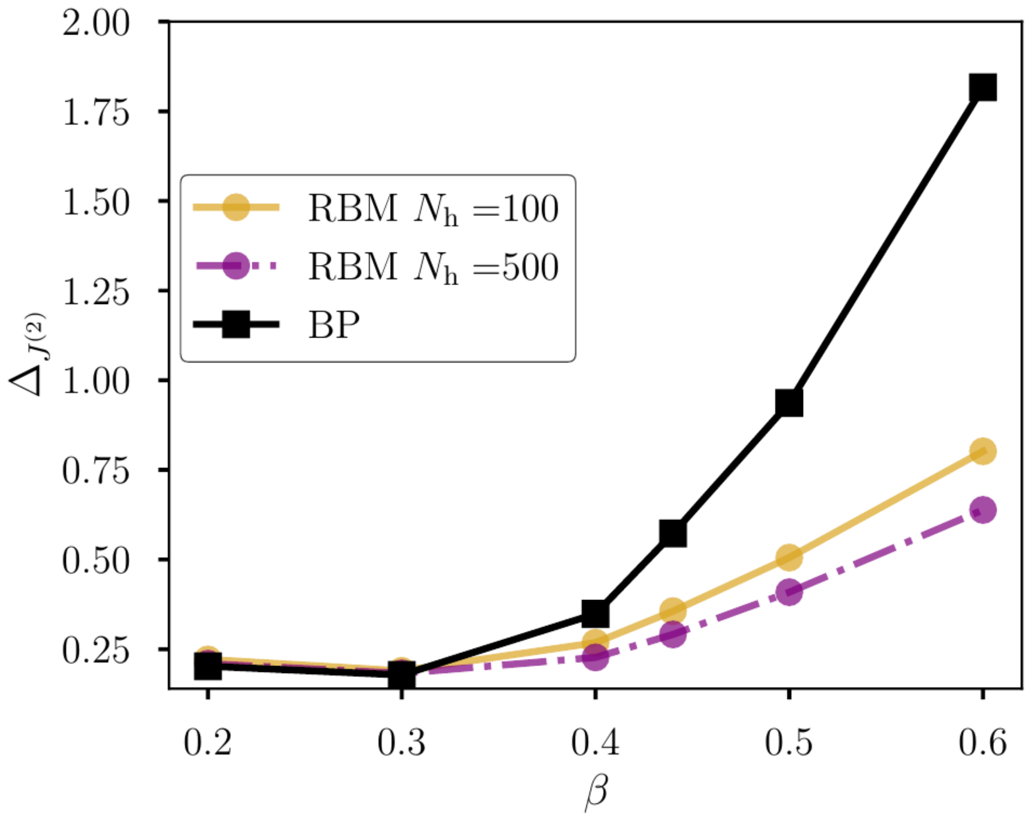
Quality comparable to standard pairwise methods

1D Ising + 3-body interactions



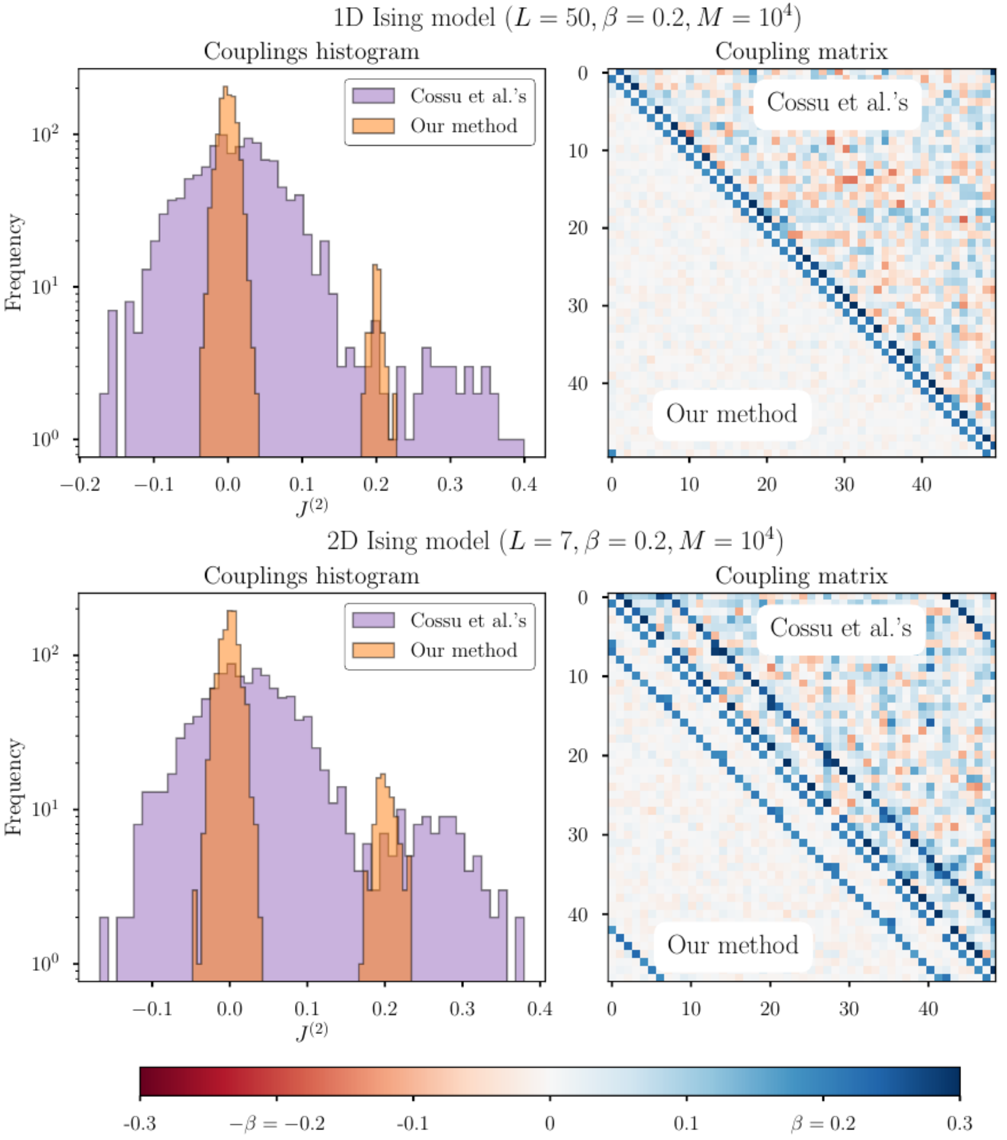
2D Ising model





Previous attempts

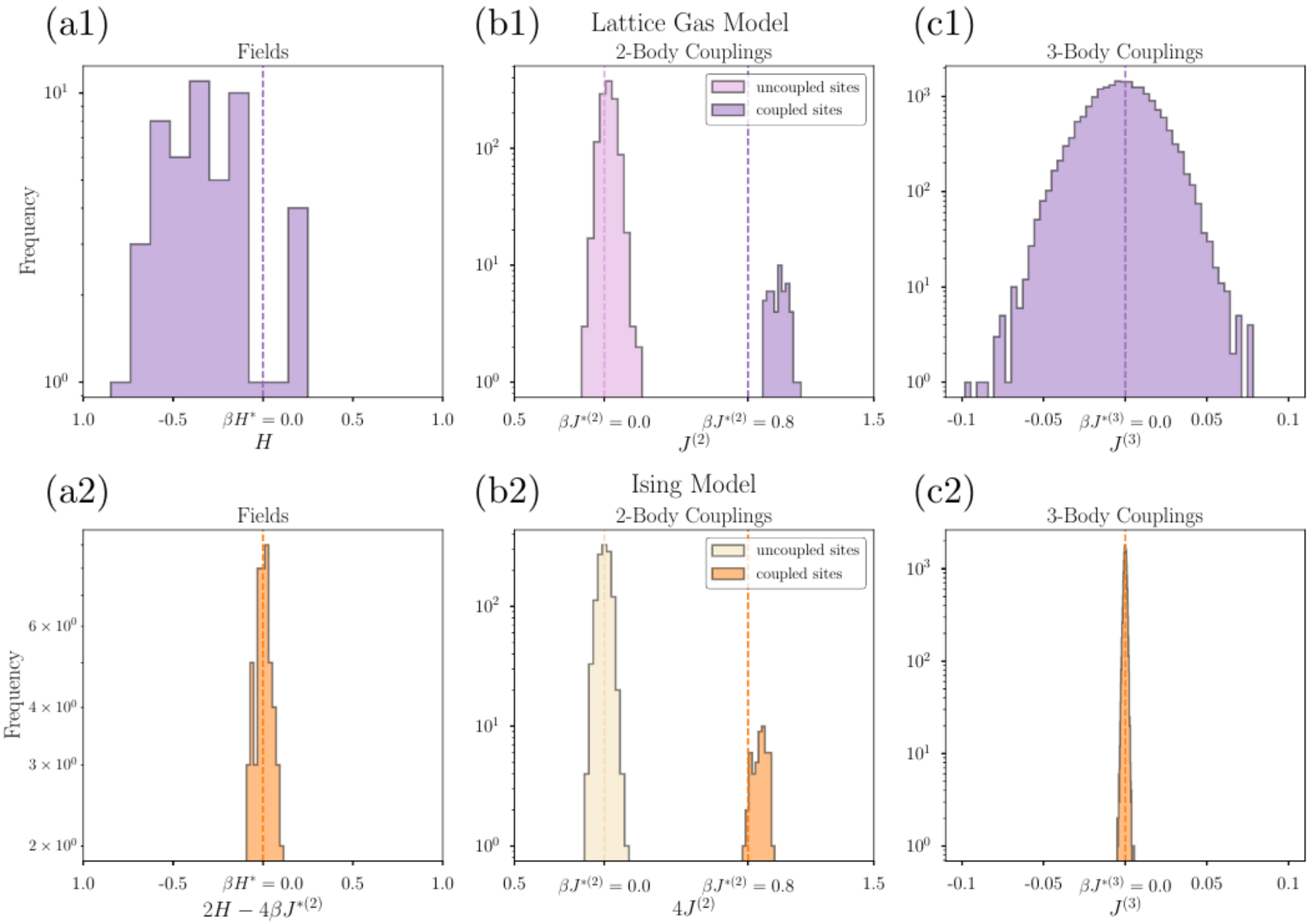
G. Cossu, L. Del Debbio, T. Giani, A. Khamseh and M. Wilson, Phys. Rev. B (2019)



Previous attempts

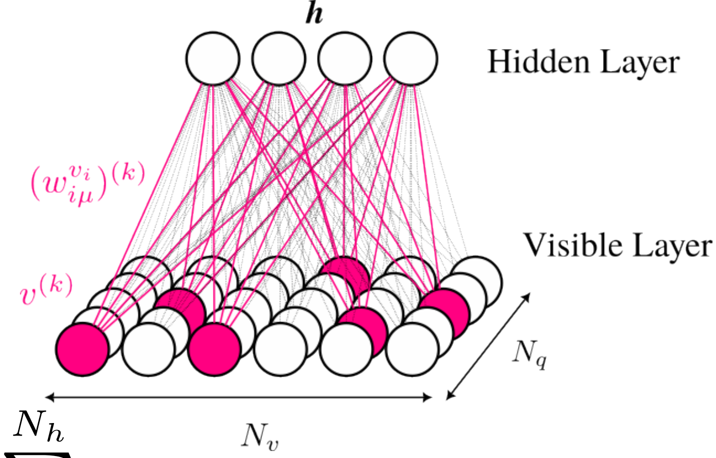
N. Bulso and Y. Roudi,
Neural Computation (2021)

Equivalence between the RBM and a lattice gas model v_i={0,1}



Beyond Ising spins

From Ising to Potts



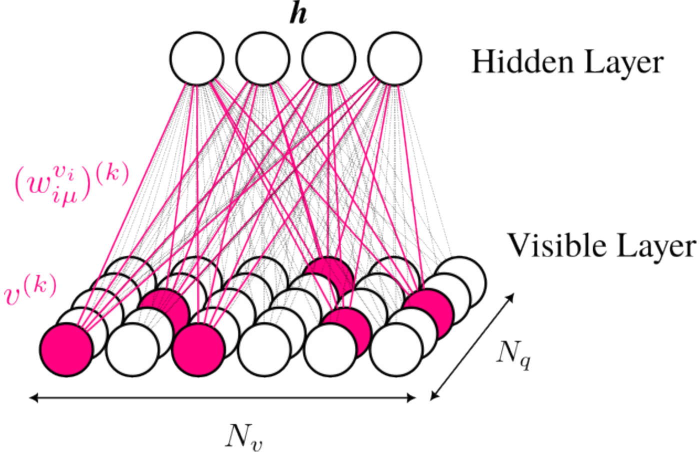
One can generalize to Potts variables

$$\mathcal{H}_{RBM}(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{N_h}\sum_{j=1}^{N_v}\sum_{a=1}^{q} h_i W_{ij}^a \delta_{a v_j} - \sum_{j=1}^{N_v}\sum_{a=1}^{q} b_j^a \delta_{a v_j} - \sum_{i=1}^{N_h} c_i h_i .$$

$$\mathcal{H}_{\mathrm{RBM}}(\boldsymbol{v}) = -\sum_j \sum_a b_j^a \delta_{a v_j} - \sum_i \ln \sum_{h_i} \exp\left(c_i h_i + h_i \sum_j \sum_a W_{ij}^a \delta_{a v_j} \right)$$

$$= -\sum_i \kappa_i^{(0)} - \sum_j \sum_a \left(b_j^a + \sum_i \kappa_i^{(1)} W_{ij}^a \right) \delta_{a v_j} - \sum_{k>1} \frac{1}{k!} \sum_{j_1,\ldots,j_k} \sum_{a_1,\ldots,a_k} \left(\sum_i \kappa_i^{(k)} W_{ij_1}^{a_1} \cdots W_{ij_k}^{a_k} \right) \delta_{a_1 v_{j_1}} \cdots \delta_{a_k v_{j_k}}$$

From Ising to Potts



We can use it to infer

$$J^{q_1,\cdots q_n}_{i_1\cdots i_n}(\boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta})$$

$$\mathcal{H}_{\mathrm{RBM}}(\boldsymbol{v}) = -\sum_j \sum_a b^a_j \delta_{av_j} - \sum_i \ln \sum_{h_i} \exp\left(c_i h_i + h_i \sum_j \sum_a W^a_{ij} \delta_{av_j}\right)$$

$$= -\sum_i \kappa_i^{(0)} - \sum_j \sum_a \left(b^a_j + \sum_i \kappa_i^{(1)} W^a_{ij}\right) \delta_{av_j} - \sum_{k>1} \frac{1}{k!} \sum_{j_1,\ldots,j_k} \sum_{a_1,\ldots,a_k} \left(\sum_i \kappa_i^{(k)} W^{a_1}_{ij_1} \cdots W^{a_k}_{ij_k}\right) \delta_{a_1 v_{j_1}} \cdots \delta_{a_k v_{j_k}}$$

Main difficulty: gauge symmetry

$$\mathcal{H}_{RBM}(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{N_h}\sum_{j=1}^{N_v}\sum_{a=1}^{q} h_i W_{ij}^a \delta_{av_j} - \sum_{j=1}^{N_v}\sum_{a=1}^{q} b_j^a \delta_{av_j} - \sum_{i=1}^{N_h} c_i h_i .$$

Invariant under the transformation

$$W_{ij}^a \to W_{ij}^a + A_{ij}$$
$$b_j^a \to b_j^a + B_j$$
$$c_i \to c_i - \sum_j A_{ij}$$

The gauge transformation changes all orders of interaction !

And the zero sum gauge in the RBM is not equivalent to the zero sum gauge in the effective Potts model

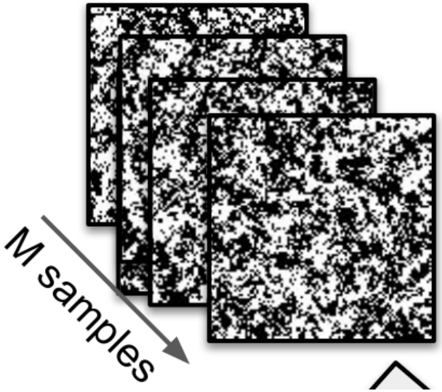
Blume-Emery-Griffiths Model model

Model for liquid ^{4}He–^{3}He mixtures,

$$\mathcal{H}_{\mathrm{BEG}} = -J \sum_{\langle j_1 j_2 \rangle} \sigma_{j_1}\sigma_{j_2} - D\sum_j \sigma_j^2 - h\sum_j \sigma_j \qquad \sigma_j \in \{-1, 0, 1\}$$

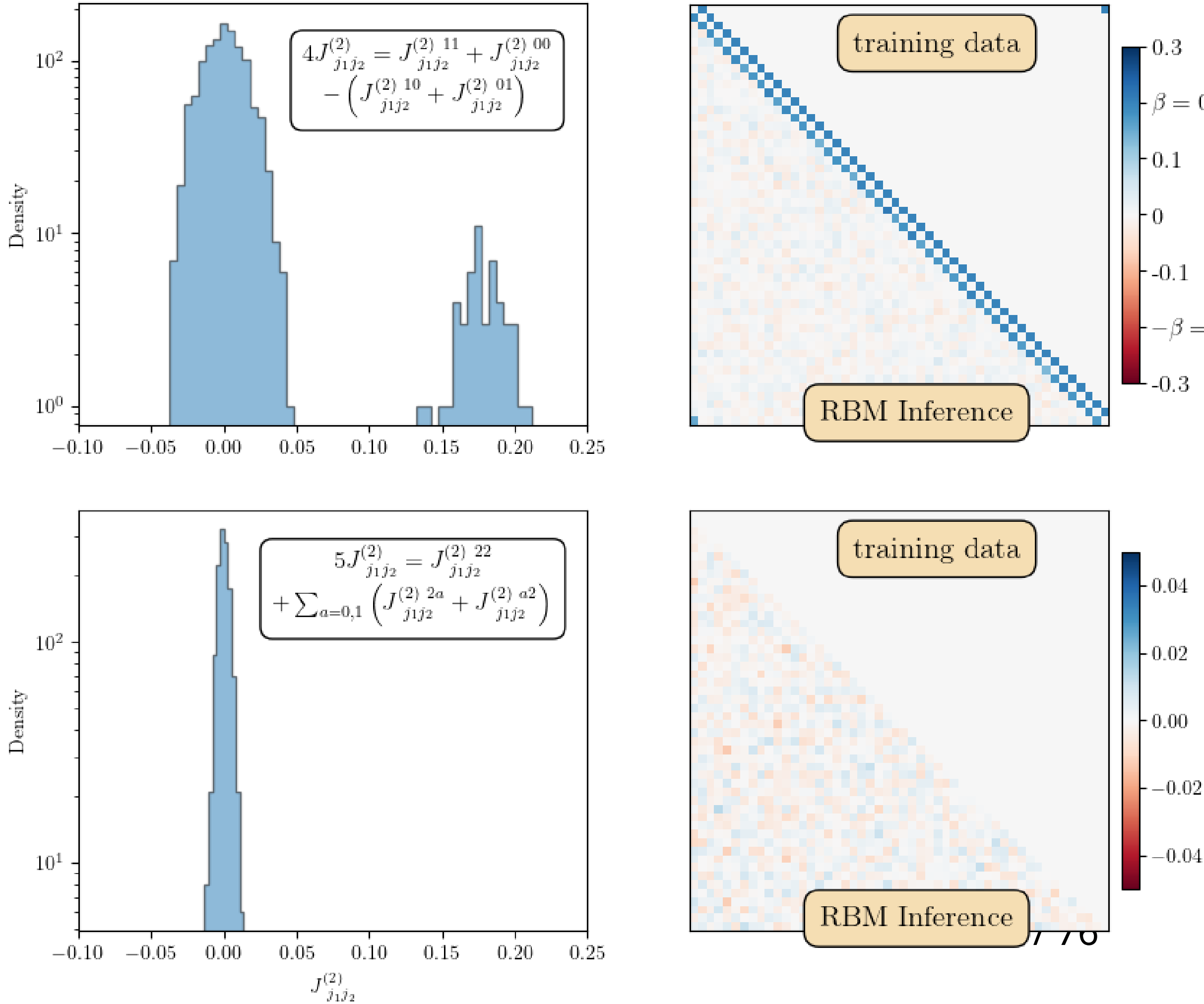
$$\mathcal{H}^{(2)}_{\mathrm{Potts}} = -\sum_{j_1<j_2}\sum_{a_1,a_2} J^{*(2)\ a_1 a_2}_{j_1 j_2}\,\delta_{a_1 v_{j_1}}\delta_{a_2 v_{j_2}} - \sum_j \sum_a H_j^{*a}\delta_{a_j}$$

$$\sigma_j = -1 \leftrightarrow v_j = 1$$
$$\sigma_j = 1 \leftrightarrow v_j = 2$$
$$\sigma_j = 0 \leftrightarrow v_j = 3,$$



Blume-Emery-Griffiths Model model

$$
J^{a_1 a_1}_{j_1 j_2} = \begin{cases} J & \text{if } a_1 = a_2 = 1 \\ J & \text{if } \ a_1 = a_2 = -1, \\ -J & \text{if } a_1 = -1, a_2 = 1 \\ -J & \text{if } \ a_1 = 1, a_2 = -1, \\ 0 & \text{if } a_1 = 0 \ \text{ or } \ a_2 = 0, \end{cases}
$$



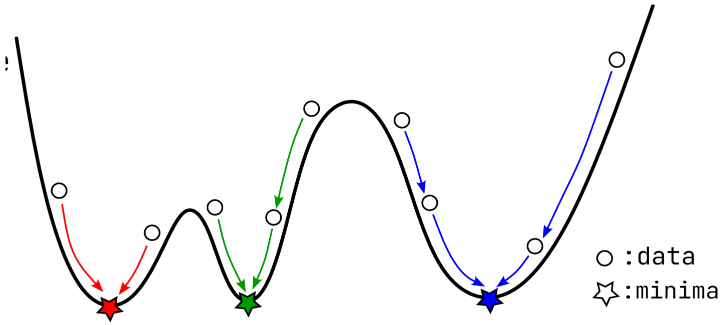
Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

Analyzing the free energy landscape

Free energy landscape

- We want to use this landscape to get a notion also to identify groups of similar sequences
- We want to obtain $f(\boldsymbol{M})$ as a function of the probability of having variables $\boldsymbol{v}$ and $\boldsymbol{h}$ activated $\boldsymbol{M}=\{\{\boldsymbol{f}_i^q\}, \{\boldsymbol{m}_a\}\}$
- $\log Z = \log \sum_{\boldsymbol{M}} e^{-Nf(\boldsymbol{M})}$ ⇒ Find the ***M***s with lower $f(\boldsymbol{M})$

We can use **basins of attraction** to cluster data points



Approximate the free energy

- We use the Plefka expansion to approximate $f(\boldsymbol{M})$
- $$f_\beta^{(2)}(\boldsymbol{M}) = f_0(\boldsymbol{M}) + \beta \left.\frac{\partial f_\beta(\boldsymbol{M})}{\partial \beta}\right|_{\beta=0} + \frac{\beta^2}{2} \left.\frac{\partial^2 f_\beta(\boldsymbol{M})}{\partial \beta^2}\right|_{\beta=0}$$

$$= \sum_{iq} f_i^q a_i^q + \sum_\mu m_\mu b_\mu - \sum_{iq} f_i^q \log f_i^q - \sum_\mu m_\mu \log m_\mu + (1-m_\mu)\log(1-m_\mu) + \beta \sum_{iq\mu} f_i^q w_{i\mu}^q m_\mu + \frac{\beta^2}{2}\sum_\mu (m_\mu - m_\mu^2) \sum_{iq} (w_{i\mu}^q)^2 f_i^q - \sum_i \sum_q w_{i\mu}^q {f_i^q}^2.$$

- Minima $\boldsymbol{\nabla} f(\boldsymbol{M}) = \boldsymbol{0}$ $\Rightarrow$ set of self-consistent equations (TAP eqs.)

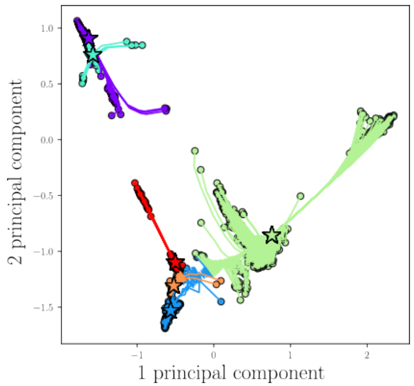
$$m_\mu[t+1] \leftarrow \mathrm{sigmoid}\left[b_\mu + \sum_{iq} f_i^q[t] w_{i\mu}^q + \left(m_\mu[t] - \frac{1}{2}\right)\left(\sum_i \left(\sum_q f_i^q[t] w_{i\mu}^q\right)^2 - \sum_{iq} (w_{i\mu}^q)^2 f_i^q[t]\right)\right]$$

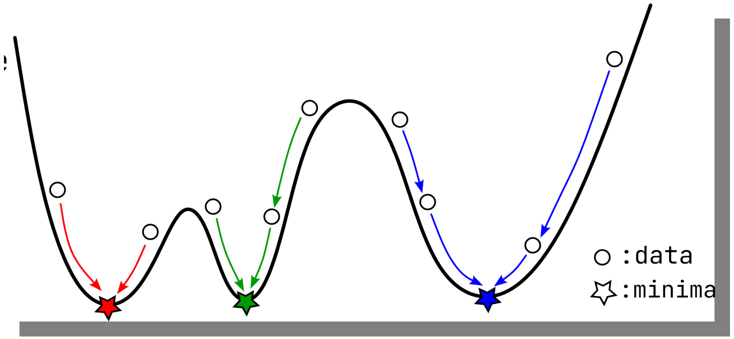
$$f_i^q[t+1] \leftarrow \mathrm{softmax}_q\left[a_i^q + \sum m_\mu[t+1] w_{i\mu}^q + \sum (m_\mu[t+1] - m_\mu^2[t+1])\left(\frac{1}{2}(w_{i\mu}^q)^2 - w_{i\mu}^q \sum_p f_i^p[t] w_{i\mu}^p\right)\right]$$

Solve iteratively

Approximate the free energy

-
-





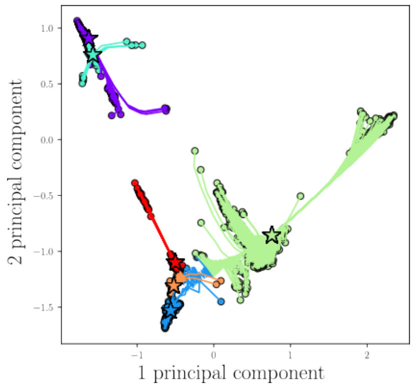
- Minima $\boldsymbol{\nabla} f(\boldsymbol{M}) = \mathbf{0}$ $\Rightarrow$ set of self-consistent equations (TAP eqs.)

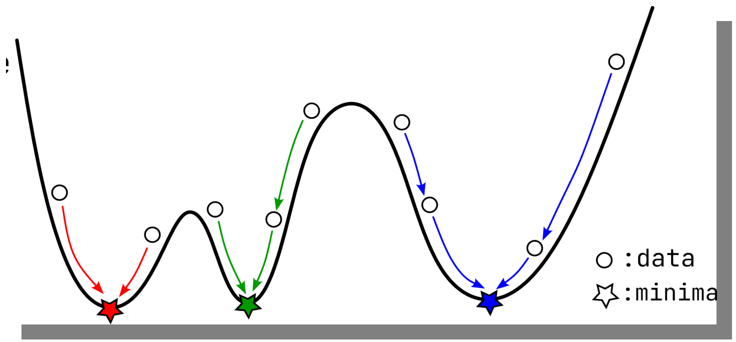
$$m_\mu[t+1] \leftarrow \mathrm{sigmoid}\left[b_\mu + \sum_{iq} f_i^q[t] w_{i\mu}^q + \left(m_\mu[t] - \frac{1}{2}\right)\left(\sum_i \left(\sum_q f_i^q[t] w_{i\mu}^q\right)^2 - \sum_{iq} (w_{i\mu}^q)^2 f_i^q[t]\right)\right]$$

$$f_i^q[t+1] \leftarrow \mathrm{softmax}_q\left[a_i^q + \sum m_\mu[t+1] w_{i\mu}^q + \sum (m_\mu[t+1] - m_\mu^2[t+1])\left(\frac{1}{2}(w_{i\mu}^q)^2 - w_{i\mu}^q \sum_p f_i^p[t] w_{i\mu}^p\right)\right]$$

Solve iteratively

Basin of attraction: class
Fixed point: "representative" features





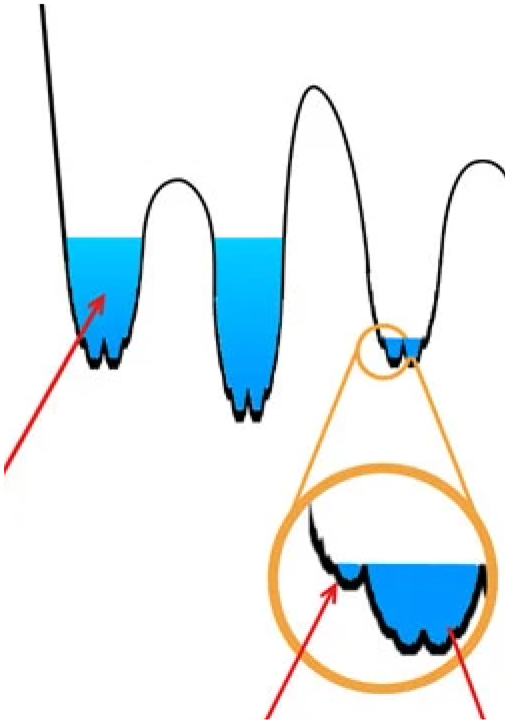
-
-
- Minima $\nabla f(\boldsymbol{M}) = \boldsymbol{0}$ ⇒ set of self-consistent equations (TAP eqs.)

$$m_\mu[t+1] \leftarrow \mathrm{sigmoid}\left[b_\mu + \sum_{iq} f_i^q[t] w_{i\mu}^q + \left(m_\mu[t] - \frac{1}{2}\right)\left(\sum_i\left(\sum_q f_i^q[t] w_{i\mu}^q\right)^2 - \sum_{iq}(w_{i\mu}^q)^2 f_i^q[t]\right)\right]$$

$$f_i^q[t+1] \leftarrow \mathrm{softmax}_q\left[a_i^q + \sum m_\mu[t+1] w_{i\mu}^q + \sum (m_\mu[t+1] - m_\mu^2[t+1])\left(\frac{1}{2}(w_{i\mu}^q)^2 - w_{i\mu}^q \sum_p f_i^p[t] w_{i\mu}^p\right)\right]$$

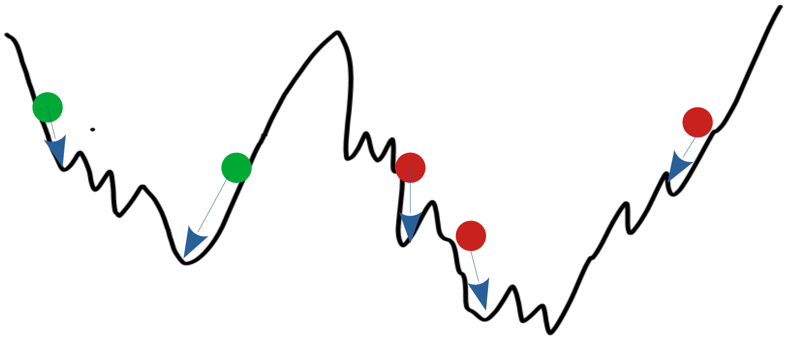
Solve iteratively

Data has a hierarchical organization

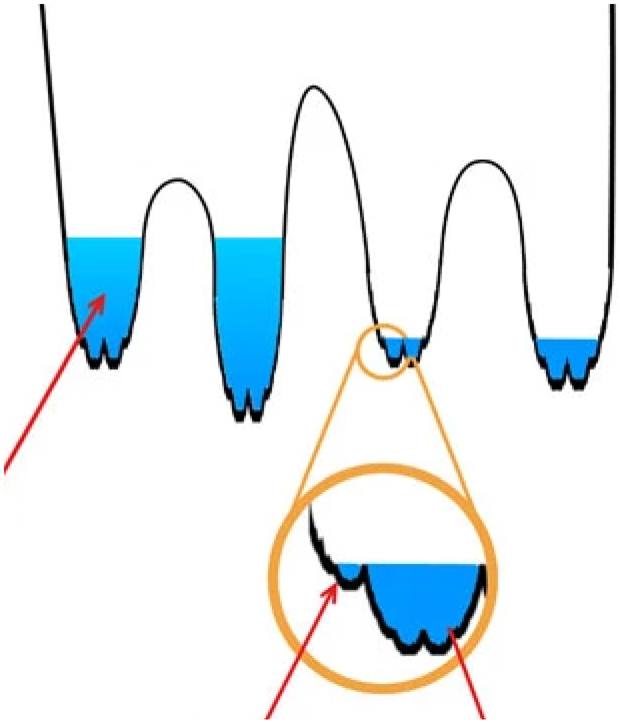


In order to be expressive enough, the RBM must describe all possible levels of similarity

The closest fixed point might be too detailed to be useful for a general classification

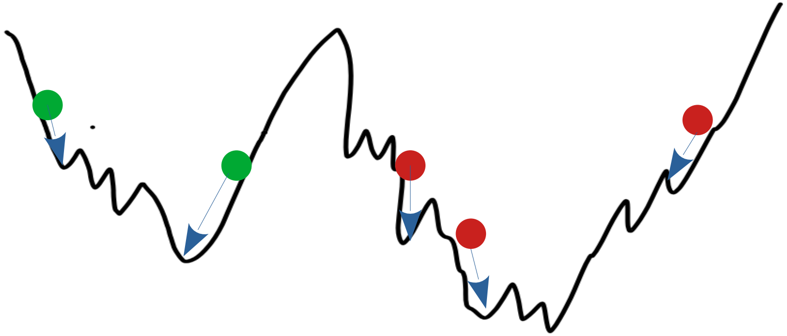


Data has a hierarchical organization



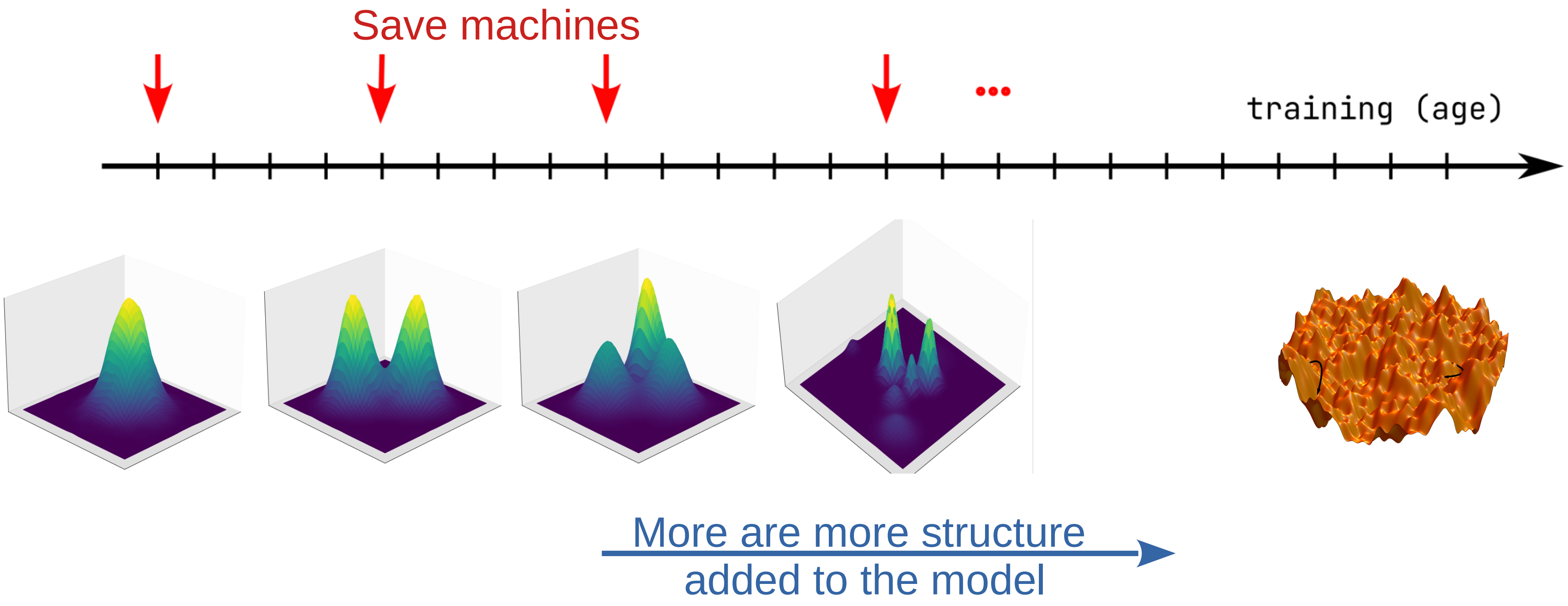
In order to be expressive enough, the RBM must describe all possible levels of similarity

The closest fixed point might be too detailed to be useful for a general classification



How do we detect larger basins?

The RBM learns in an hierarchical way

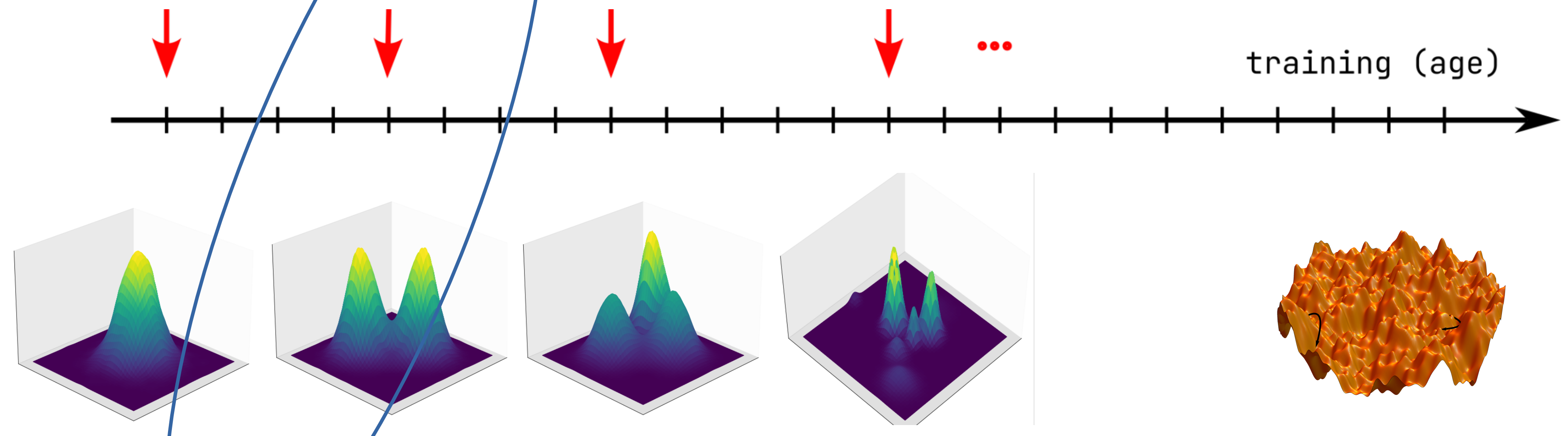


The RBM learns in an hierarchical way

The W encode the PCA
of the dataset: **Pairwise correlations**

Higher order correlations

Save machines

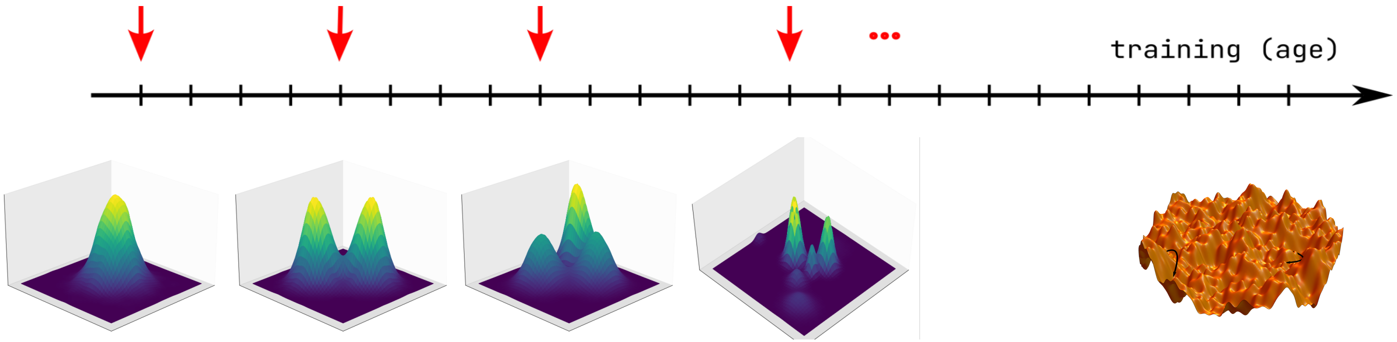


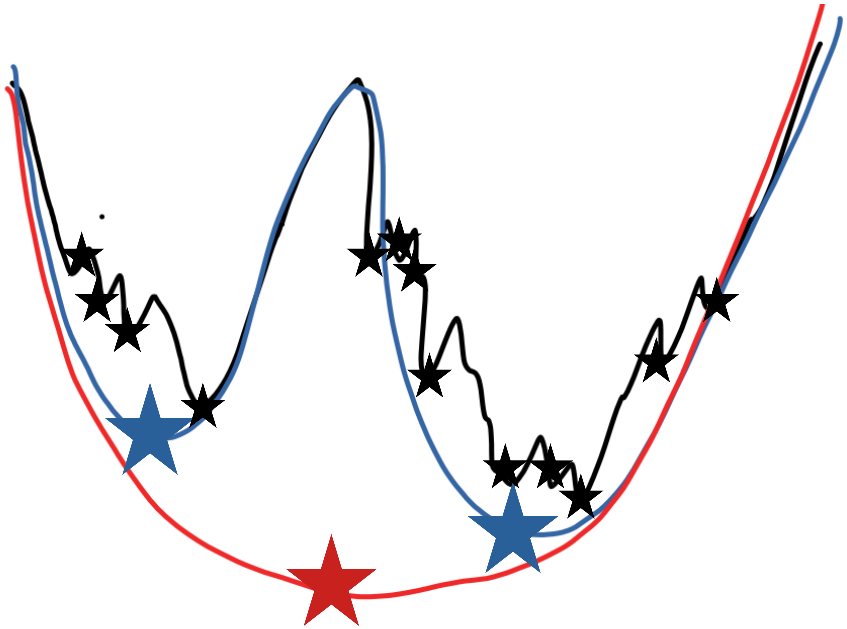
More are more structure
added to the model

* Decelle, Fissore and Furtlehner, *Spectral dynamics of learning in restricted boltzmann machines* (2017)

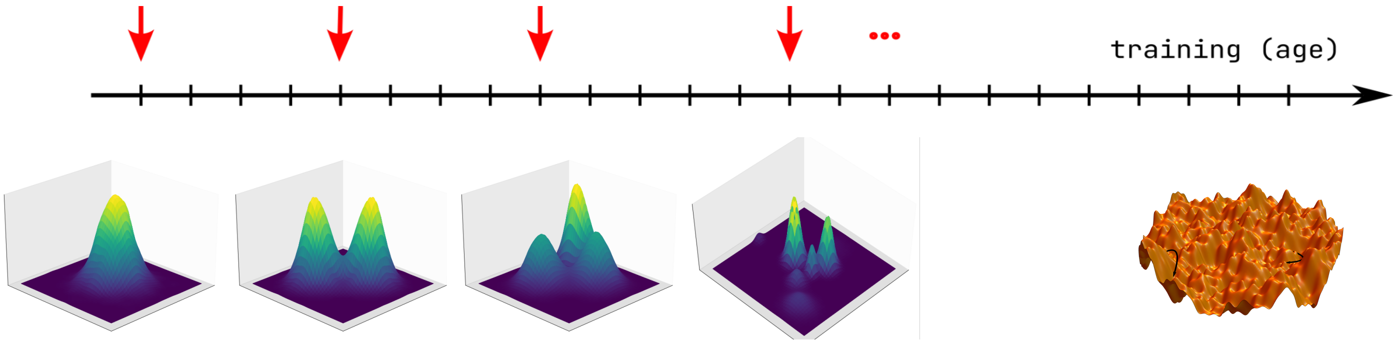
* Decelle, & Furtlehner, *Restricted Boltzmann machine: Recent advances and mean-field theory* (2021)

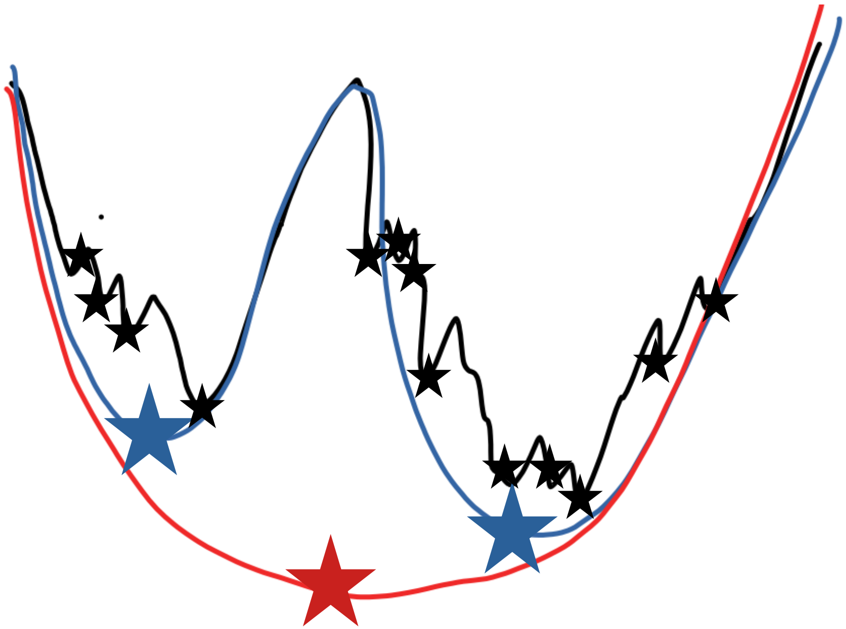
Hierarchical Clustering

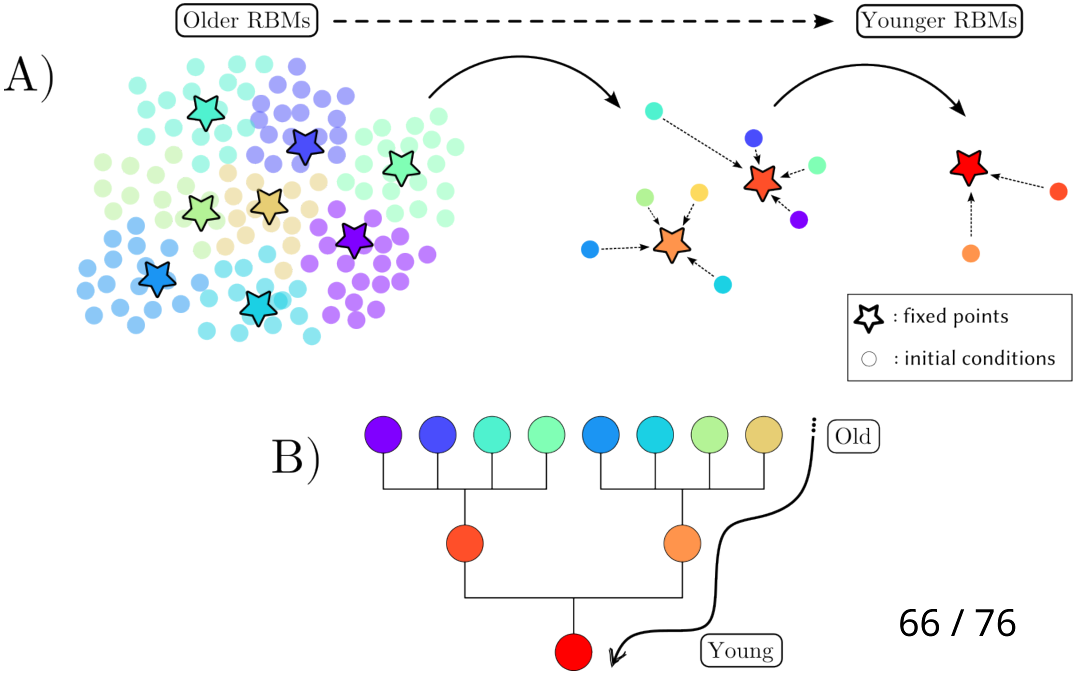




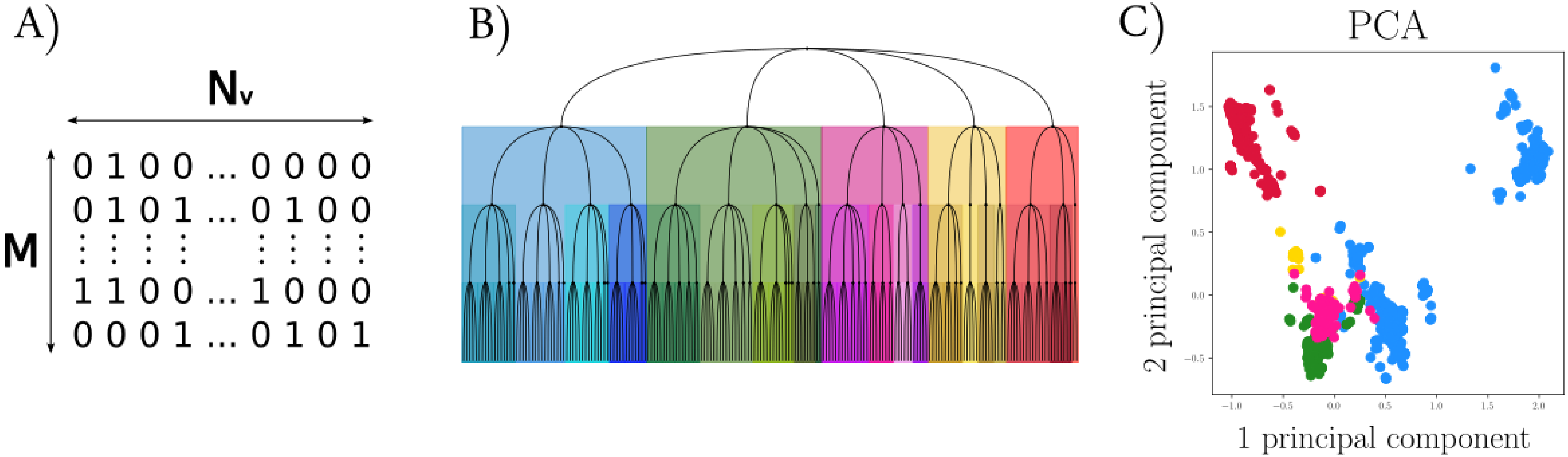
Hierarchical Clustering



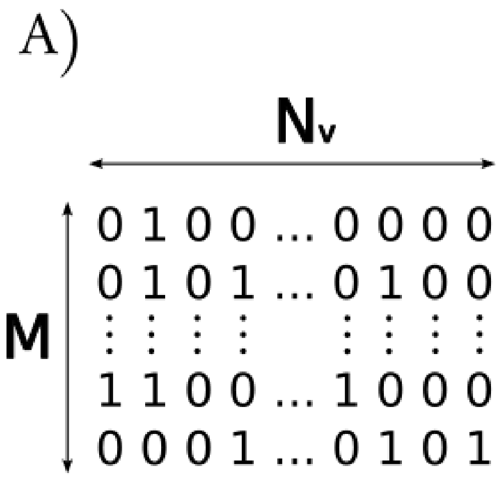


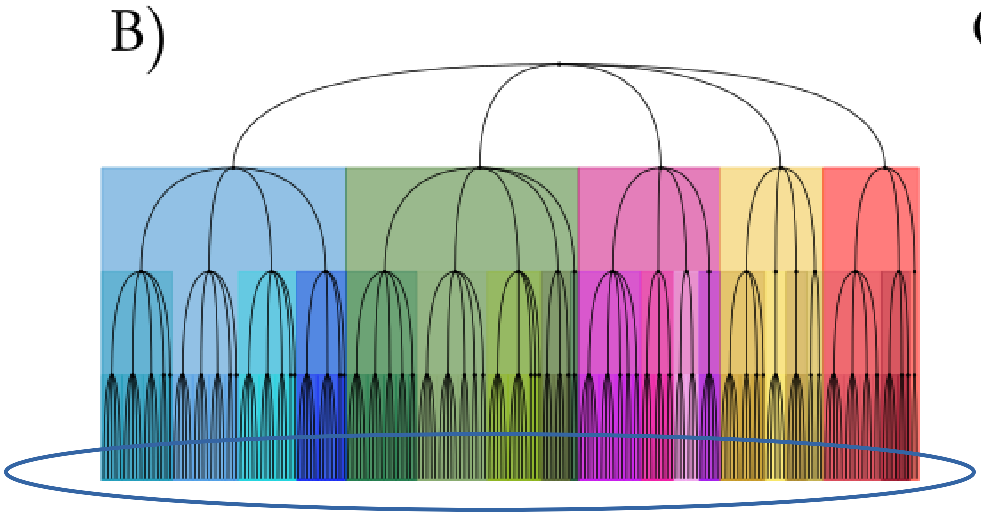


Example: synthetic evolutionary data

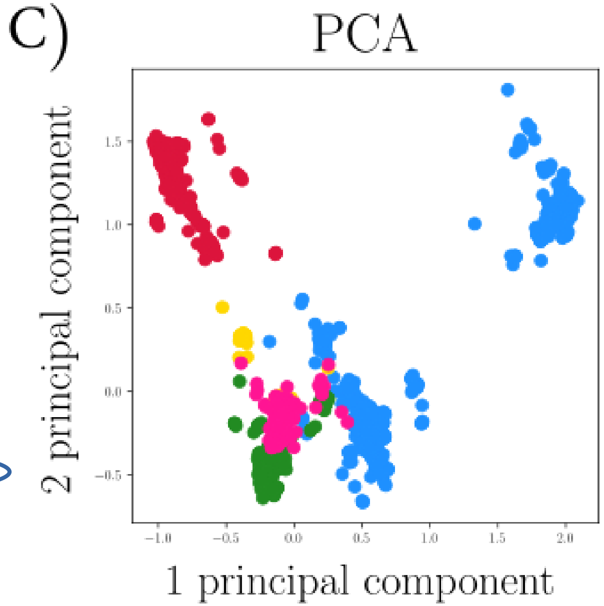


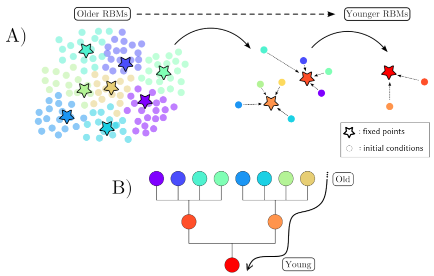
Example: synthetic evolutionary data





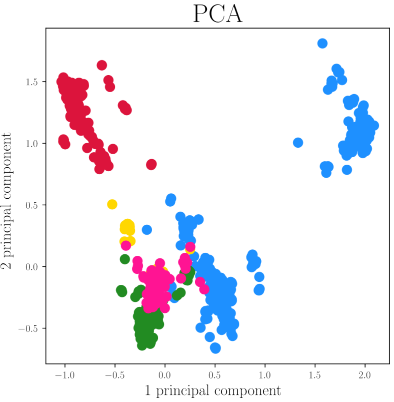
Train a RBM

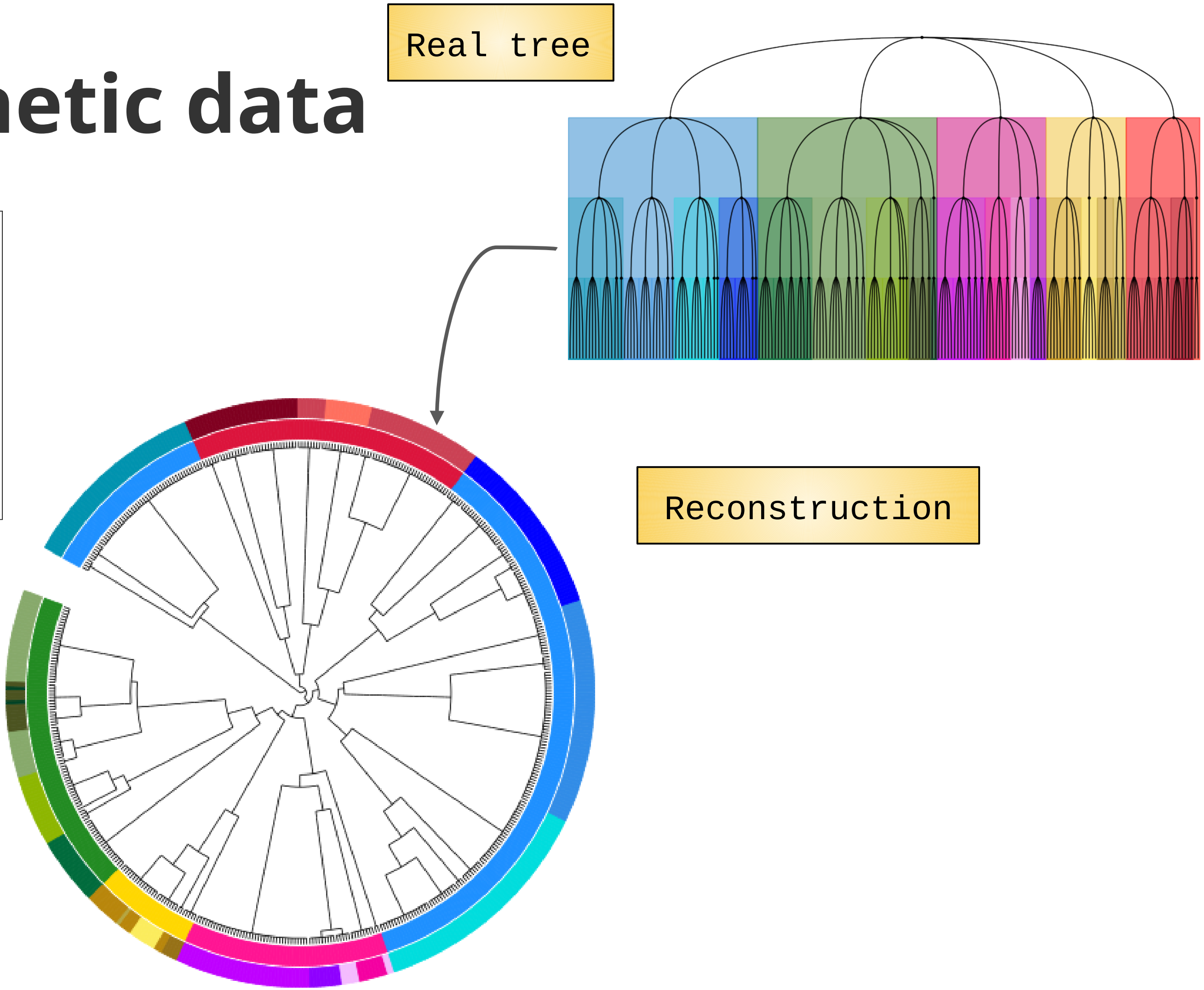




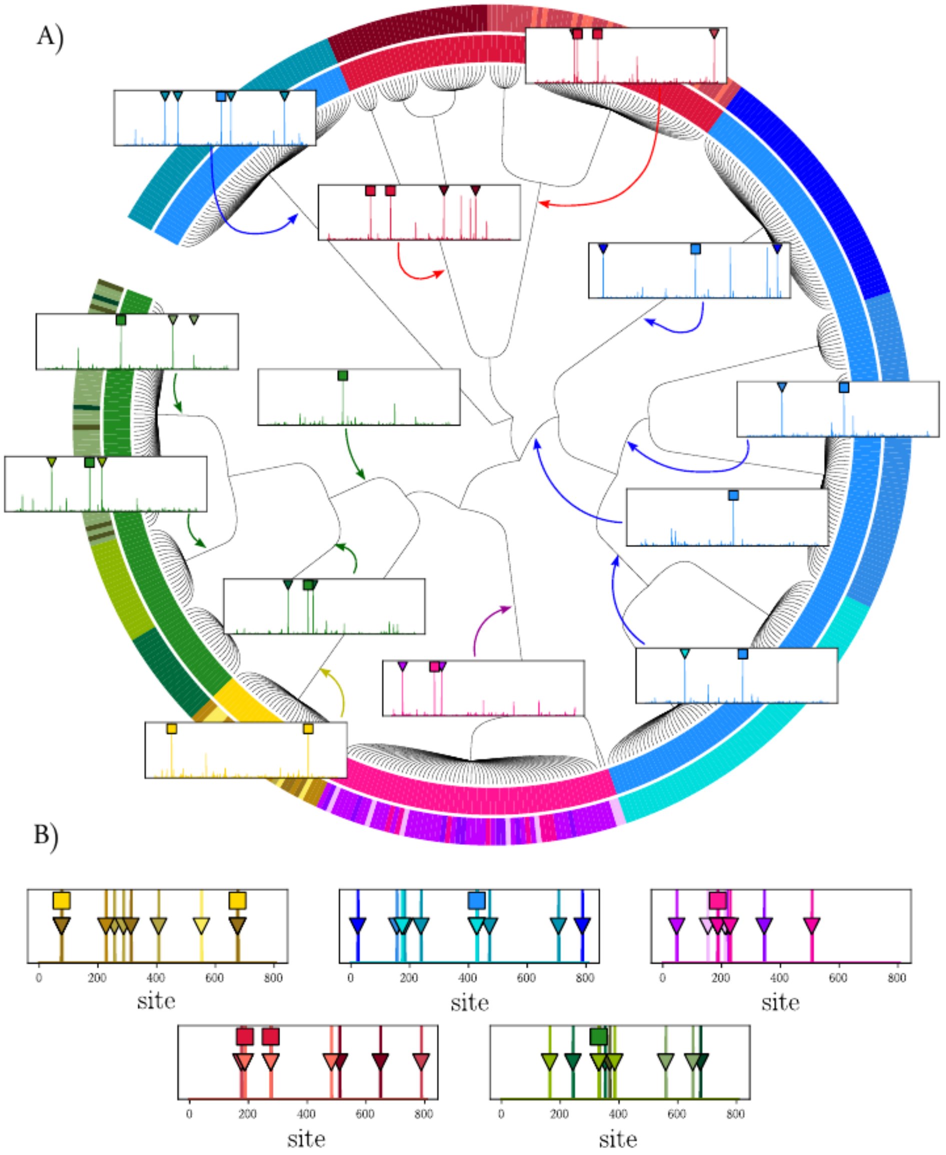
Build a tree
Using machines saved during the training

Synthetic data



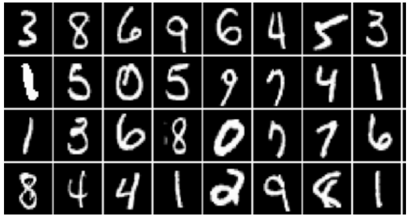


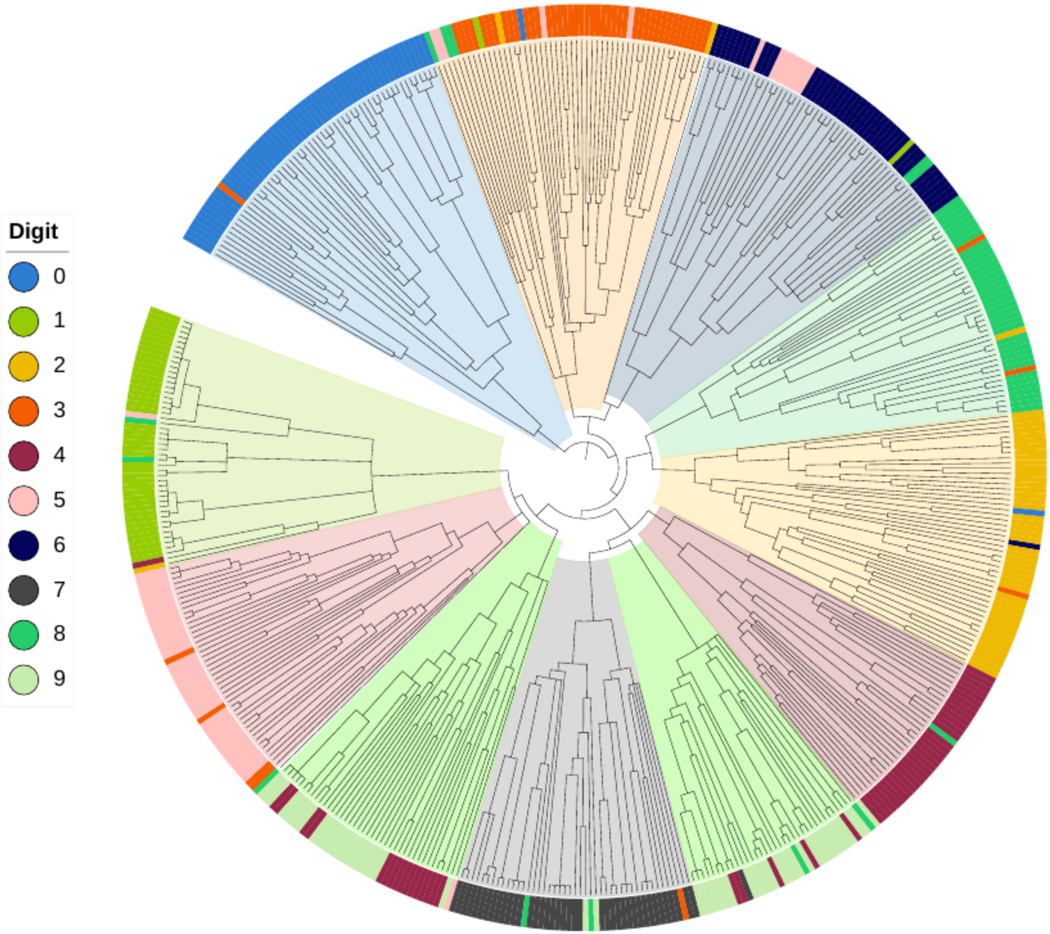
Synthetic data



Hierarchical Clustering

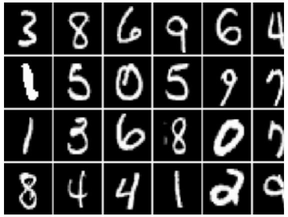
MNIST data

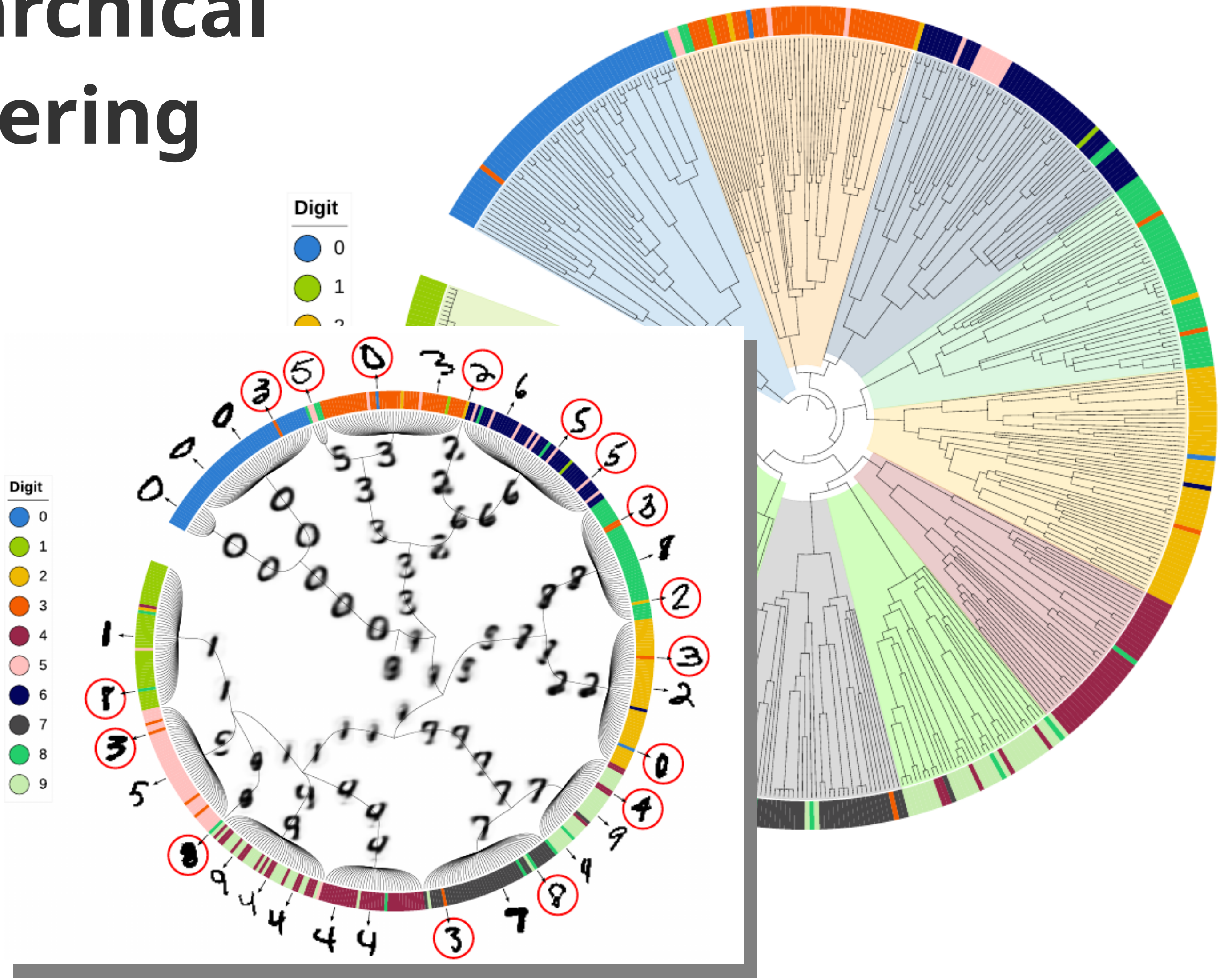




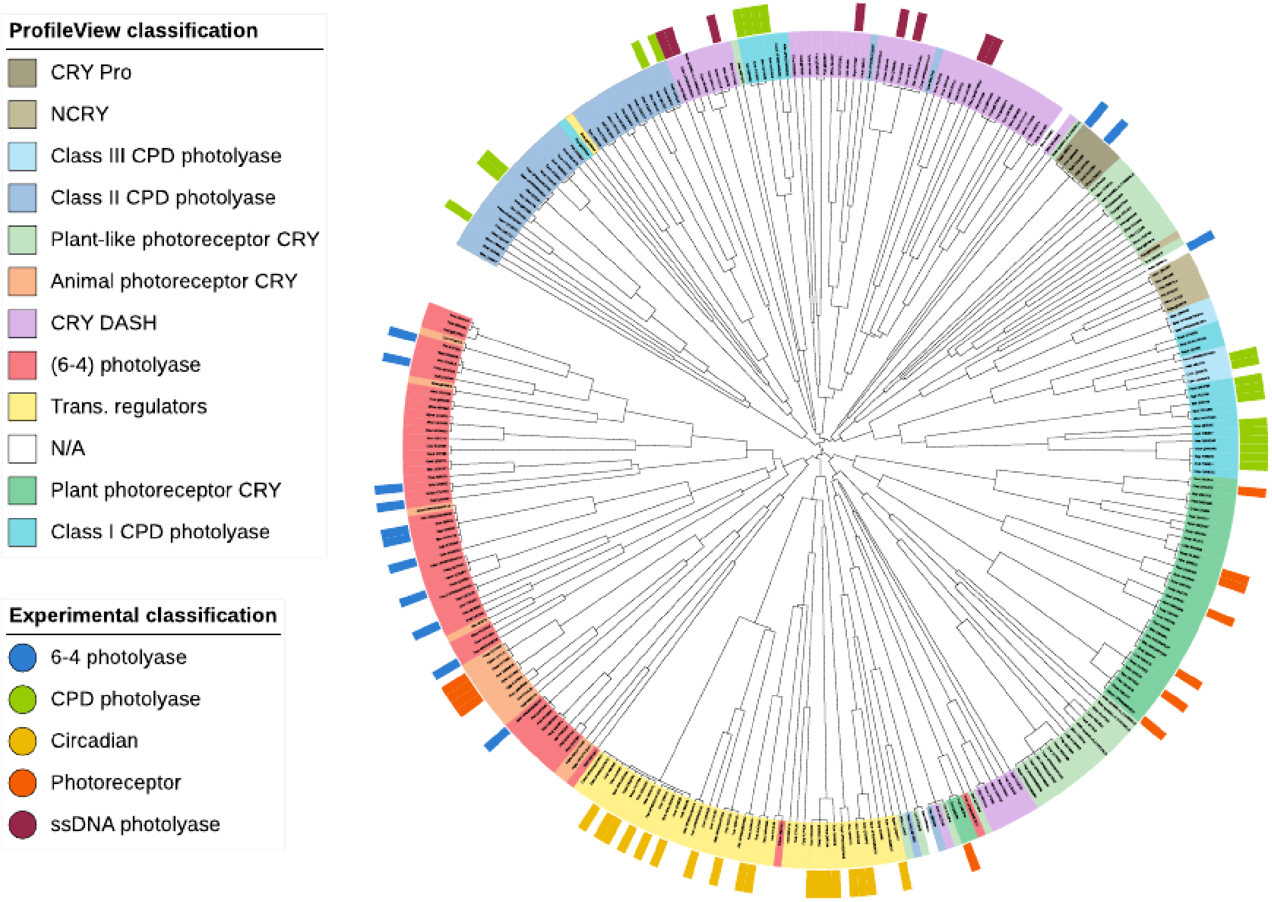
Hierarchical Clustering

MNIST data



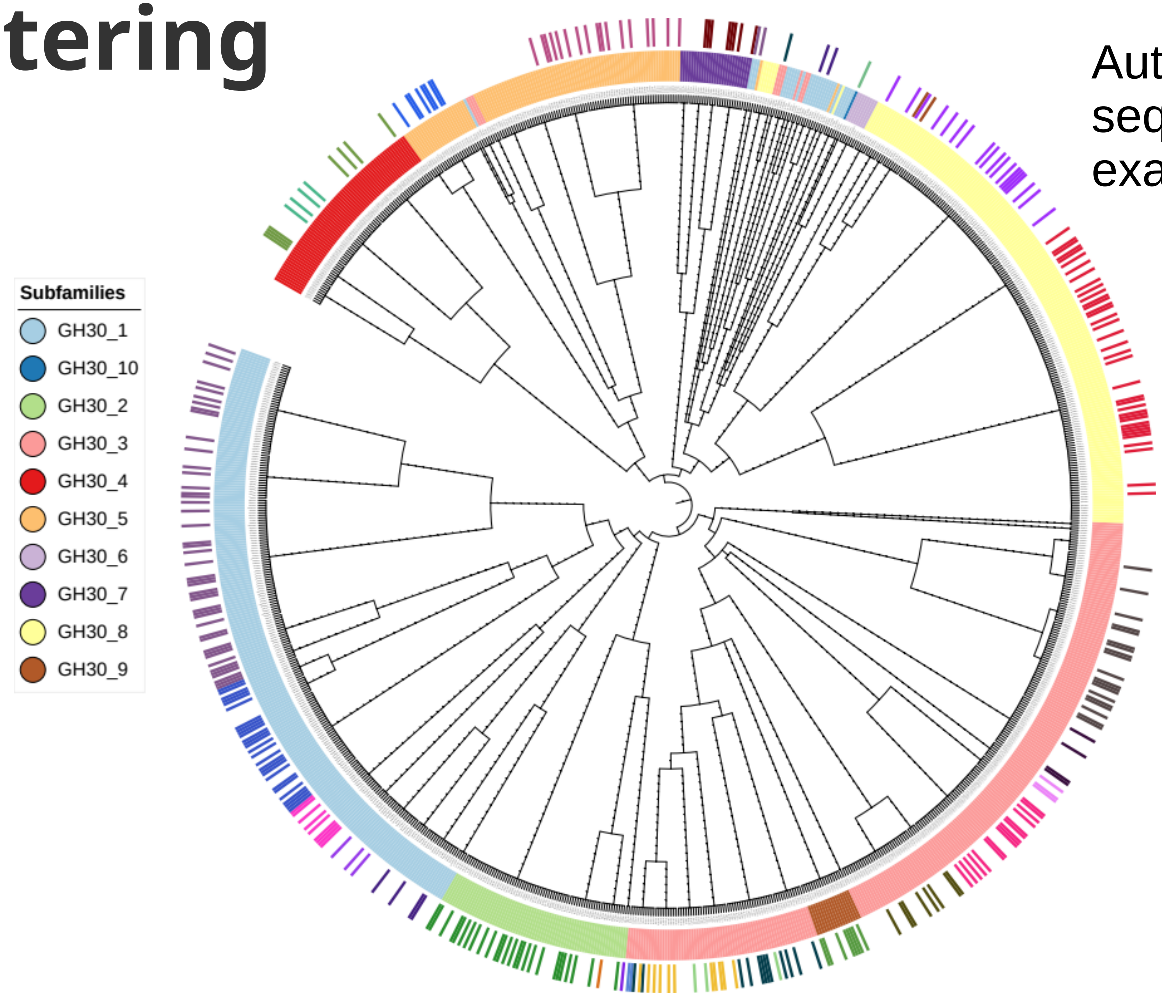


Protein function classification



CPF protein family

Hierarchical Clustering



Automatically label sequences based on a few examples

Conclusions

- RBMs are both expressive and simple
- The are as interpretable as the Boltzmann Machines
- They can be used to infer multi-body interactions without blowing the number of parameters
- We have mappings between the:
  - Bernouilli-Bernoulli RBM → Generalized Ising model
  - Bernouilli-Potts RBM → Generalized Potts model (still testing)
- We can use the RBM for hierarchical clustering

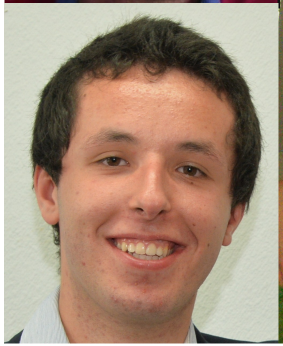
Acknowledgments

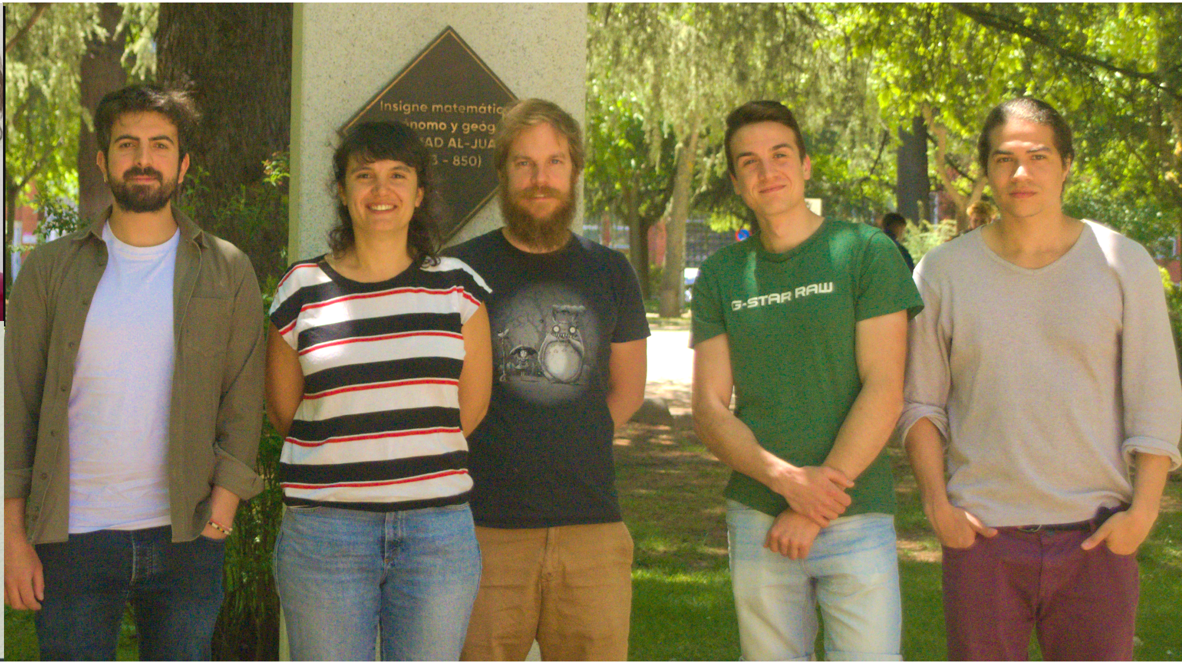
Aurélien Decelle
Giovanni Catania
Alfonso Navas
Lorenzo Rosset

UCM

Nicolas Béreux
Cyril Furtlehner

(Paris-Saclay)





Decelle, Furtlehner, Navas, Seoane, arXiv: 2309.02292 SciPost 2024

Code torchRBM
Training RBMs

Code
Inference
couplings