Class 2: Interpreting RBMs

Plan for the lecturers

- ~~Class 1: Introduction to Energy Based Models~~
- Class 2: Interpretability. How can we learn from trained networks?
- Class 3: Training optimization, the role of MCMC. How can we improve the training mechanisms by understanding their physics?

Summary

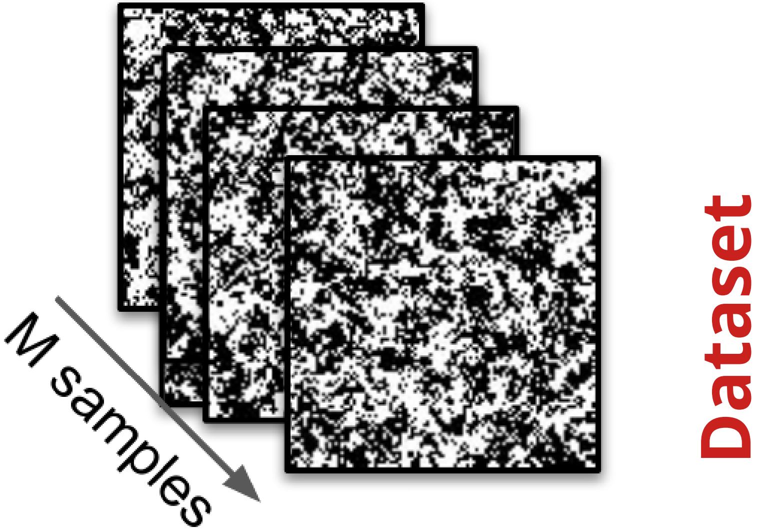
$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

- **Application 1:** Interpretation of the energy function: $E_{\theta}(x)$
  - Intro: General applications of inverse statistical mechanics
  - Mapping the RBM to a multi-body interaction Ising model
  - Inference of interaction networks
- **Application 2:** Exploring the inferred probability distribution function: $p_{\theta}(x)$
  - Probe perturbately the free-enery landscape using statistical physics
  - Use the training dynamics to reveal relational trees between data:
    - Hierarchical clustering
    - Unsupervised classification

Interpreting the **energy function**

Inverse Ising problem

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

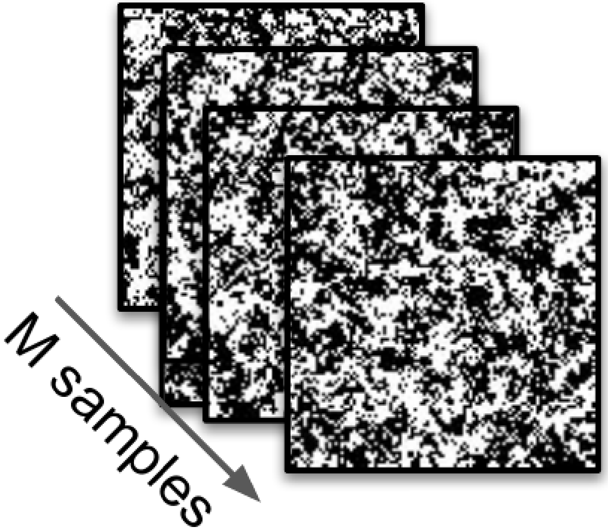


$$E_{\text{Ising 2D}}(\boldsymbol{S}) = -\hat{J} \sum_{\langle i,j \rangle} S_i S_j$$

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

Inverse Ising problem

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



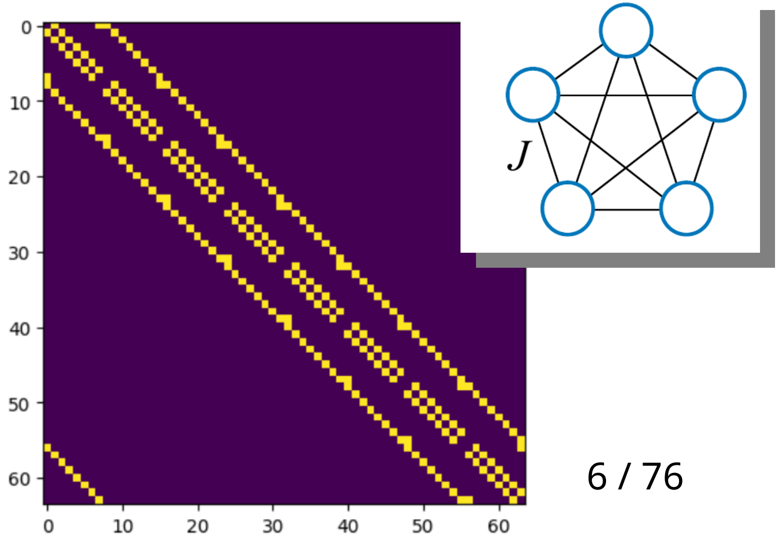
Dataset

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

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

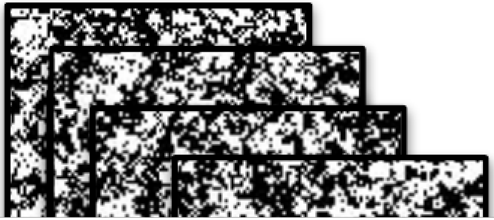
Am I able to infer which was the interaction model that generated it?

$$E_{J,h}(\boldsymbol{S}) = -\sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i$$



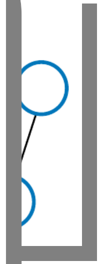
Inverse Ising problem

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



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

$$E_{J,h}(\boldsymbol{S}) = -\sum J_{ij} S_i S_j - \sum h_i S_i$$



$$p_{\text{data}}(\boldsymbol{S}) = \frac{1}{Z} e^{\beta \hat{J} \sum_{\langle i,j \rangle} S_i S_j}$$

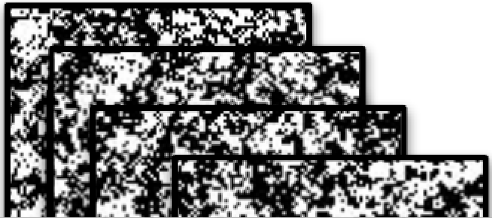
$$\beta \hat{J}_{ij} = J_{ij} \quad h_i = 0$$

Solution is unique !

$$p_{J,h}(\boldsymbol{S}) = \frac{1}{Z} e^{\sum_{ij} J_{ij} S_i + \sum_i h_i S_j}$$

Inverse Ising problem

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



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

$$E_{J,h}(\boldsymbol{S}) = -\sum J_{ij} S_i S_j - \sum h_i S_i$$

$$p_{\text{data}}(\boldsymbol{S}) = \frac{1}{Z} e^{\beta \hat{J} \sum_{\langle i,j \rangle} S_i S_j}$$

$$p_{J,h}(\boldsymbol{S}) = \frac{1}{Z} e^{\sum_{ij} J_{ij} S_i + \sum_i h_i S_j}$$

$$\beta \hat{J}_{ij} = J_{ij} \qquad h_i = 0$$

Solution is unique !

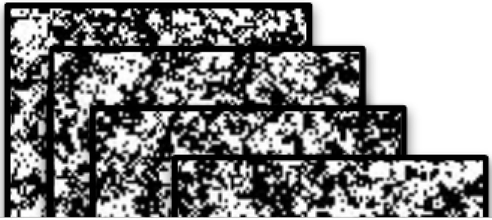
Fixed point

$$\langle S_i S_j \rangle_{p_{J,h}} = \langle S_i S_j \rangle_{p_{\text{data}}}$$

$$\langle S_i \rangle_{p_{J,h}} = \langle S_i \rangle_{p_{\text{data}}}$$

Inverse Ising problem

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



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

$$E_{J,h}(\boldsymbol{S}) = -\sum J_{ij} S_i S_j - \sum h_i S_i$$

$$p_{\text{data}}(\boldsymbol{S}) = \frac{1}{Z} e^{\beta \hat{J} \sum_{\langle i,j \rangle} S_i S_j}$$

We only Know the data

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} \delta\left(\boldsymbol{x} - \boldsymbol{x}^{(m)}\right)$$

$$p_{J,h}(\boldsymbol{S}) = \frac{1}{Z} e^{\sum_{ij} J_{ij} S_i + \sum_i h_i S_j}$$

$$\beta \hat{J}_{ij} \neq J_{ij} \qquad h_i \neq 0$$

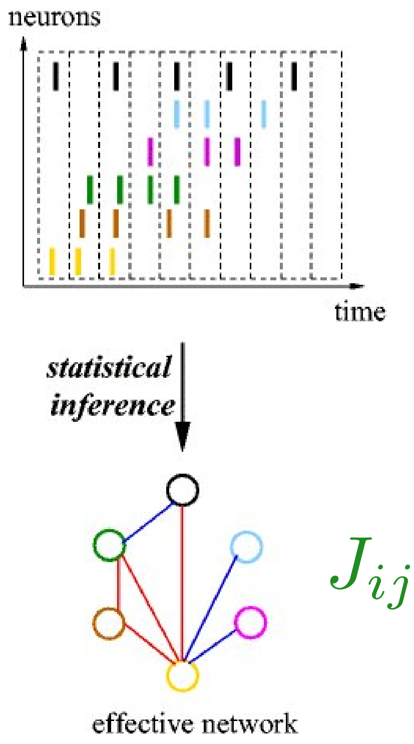
Solution is unique !

Fixed point

$$\langle S_i S_j \rangle_{p_{J,h}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}}$$

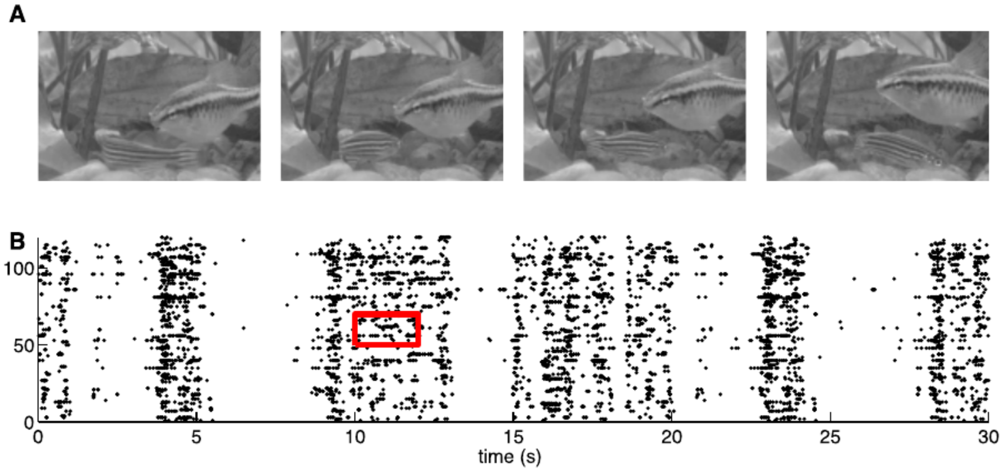
$$\langle S_i \rangle_{p_{J,h}} = \langle S_i \rangle_{p_{\mathcal{D}}}$$

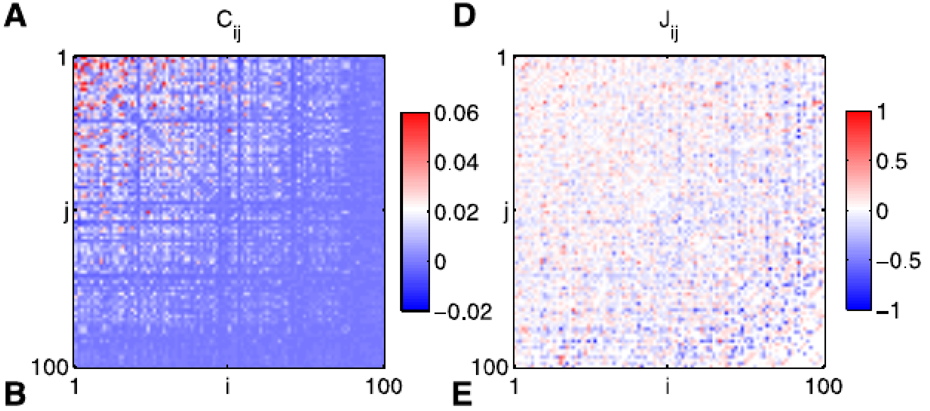
Applications I: reconstruction of neural connections



J_{ij}

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



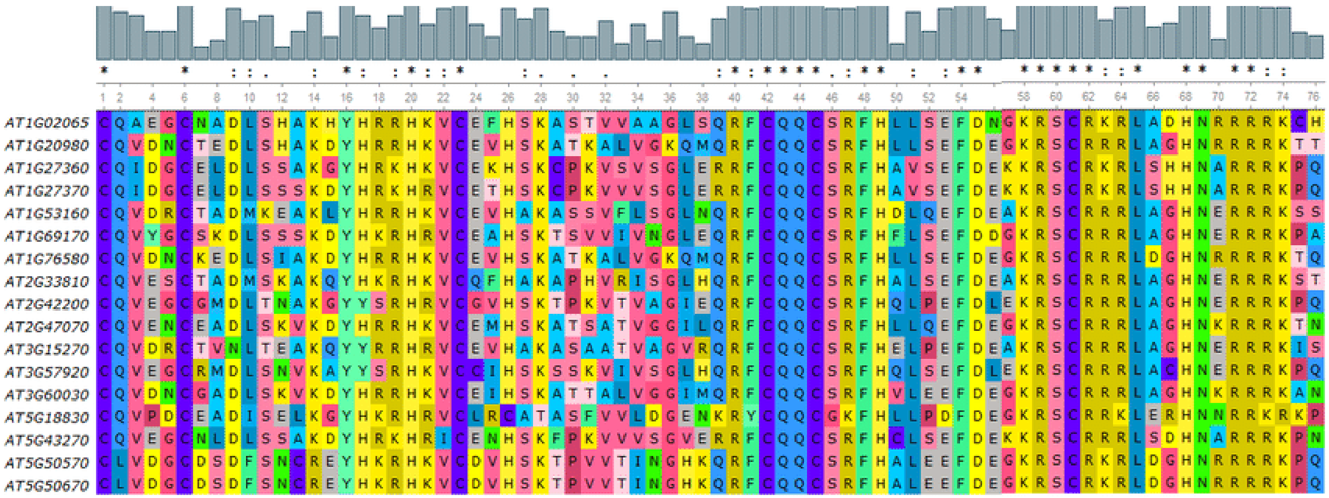


Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., & Berry, M. J. (2014).

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)

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



MSA

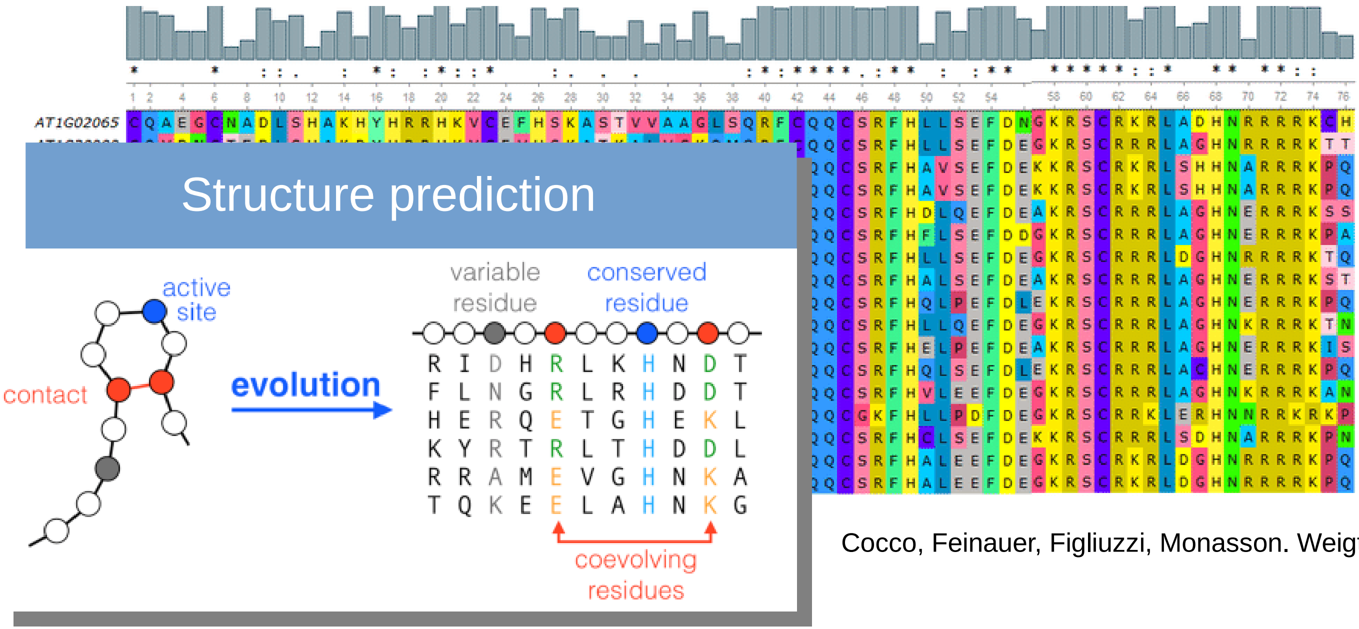
q=21

Model the “true” *fitness landscape*

Statistical sequence landscape

Applications II: Inverse Potts
Direct coupling analysis (DCA)

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



MSA

q=21

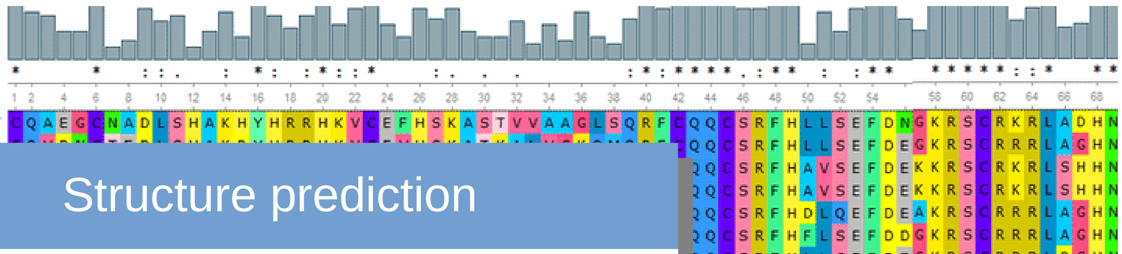
Model the “true”
fitness landscape

Statistical sequence landscape

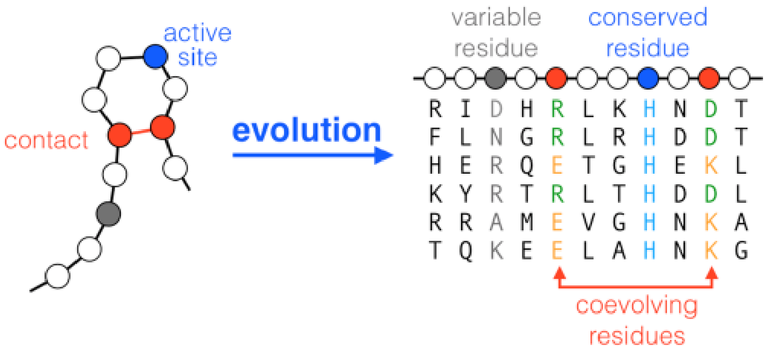
Cocco, Feinauer, Figliuzzi, Monasson. Weigt, Rep. Prog. Phys. 81 (2018) 032601

Ex. Inverse Potts
Direct coupling analysis (DCA)

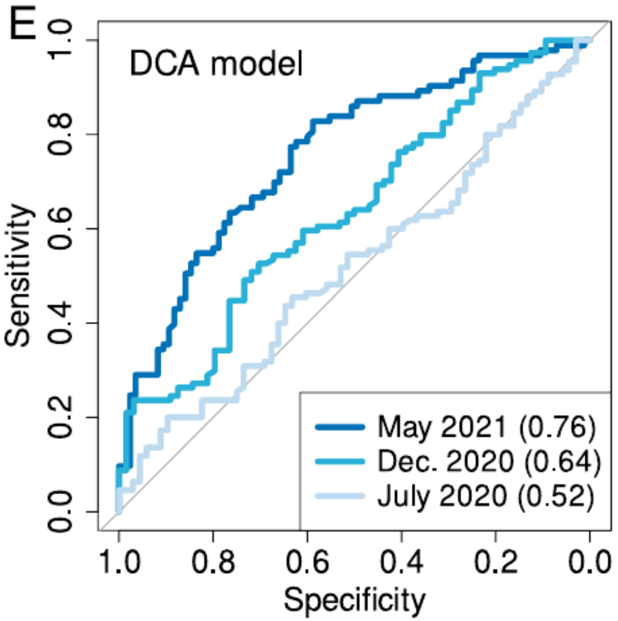
$$E_{J,h}(\boldsymbol{x}) = -\sum_{i,j=1}^{N_v}\sum_{q_1,_2=1}^{N_q} J_{ij}^{q_1,q_2}\,\delta_{x_i,q_1}\delta_{S_i,q_2} - \sum_{i=1}^{N_v}\sum \cdots q\}$$



Structure prediction



Mutation prediction

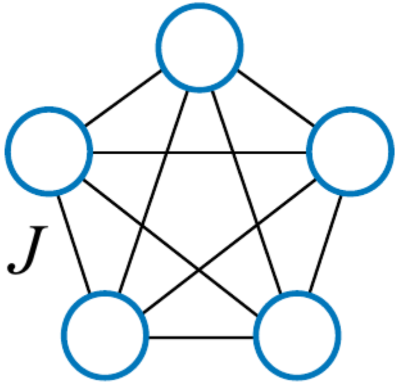


Cocco, Feinauer, Figliuzzi, Monasson. Weigt, Rep. Prog. Phys. 81 (2018) 032601

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

Pairwise models : The Boltzmann machine

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



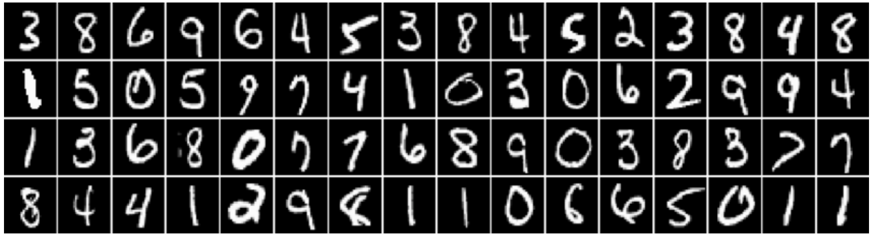
Simple and easy to interpret, but are strongly limited...

Pairwise models : The Boltzmann machine

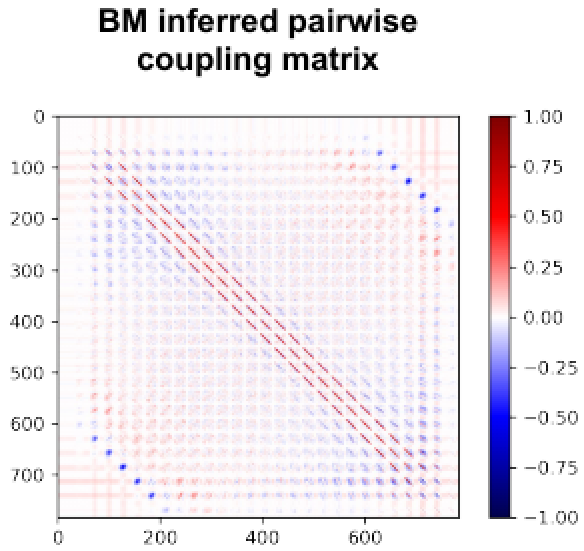
Hinton and Sejnowski (1983)

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

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



learning



Pairwise models : The Bolt

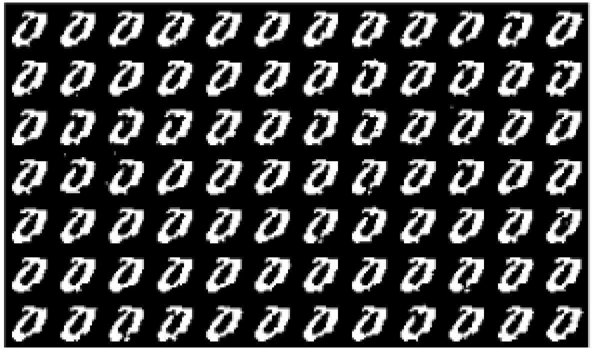
We need to encode **higher order correlations** !

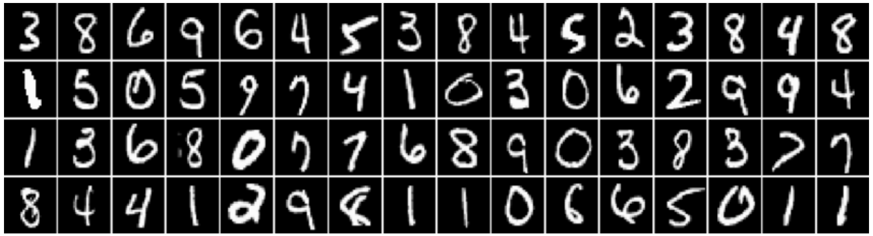
$$E_{J,h}(\boldsymbol{x}) = -\sum_{ij} J_{ij} x_i x_j - \sum_i h_i x_i$$

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

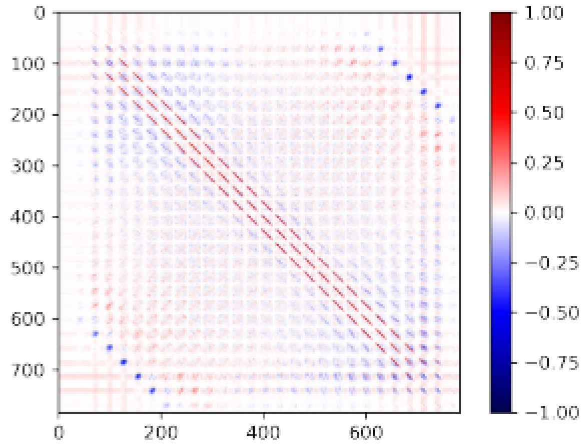
Generation

Samples generated with the BM

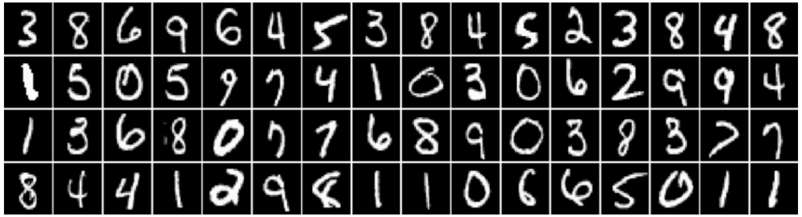


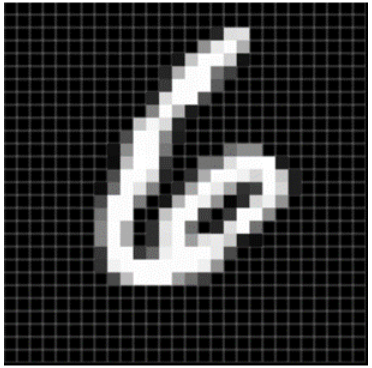


learning



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

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

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

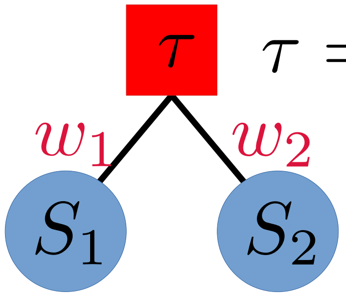
$$f_{ijk} = \langle x_i x_j x_k \rangle_{\text{data}}$$

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

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



$\tau = \pm 1$

$S_i = \pm 1$

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

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

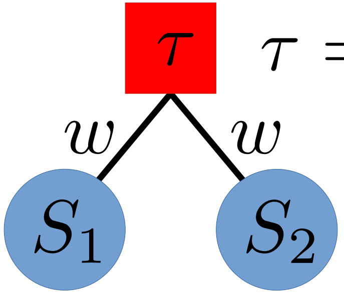
$$\mathcal{H} = -\log \sum_{\tau=\pm 1} e^{\tau(w_1 S_1 + w_2 S_2)} = -\log 2 \cosh\left[w_1 S_1 + w_2 S_2\right]$$

$$= -J S_1 S_2 - J$$

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



$\tau = \pm 1$

$S_i = \pm 1$

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

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

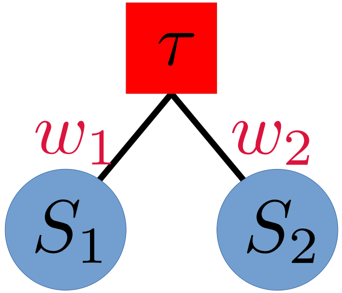
$$\mathcal{H} = -\log \sum_{\tau=\pm 1} e^{w\tau(S_1+S_2)} = -\log 2\cosh\left[w(S_1 + S_2)\right]$$

$$= -JS_1S_2 - J$$

$$S^{2k} = 1$$

$$\Rightarrow \boxed{\cosh 2w = e^{2J}} \qquad J > 0 \qquad S^{2k+1} = S$$

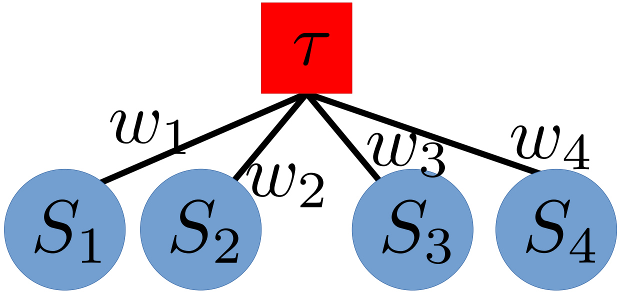
Alternative solution: add hidden variables



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

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

Alternative solution: add hidden variables



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

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

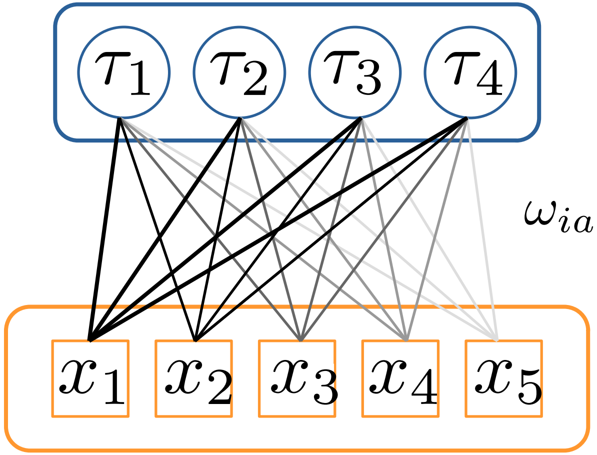
$$= -J^{(4)}_{1234} S_1 S_2 S_3 S_4 - J^{(2)}_{12} S_1 S_2 - J^{(2)}_{13} S_1 S_3 - J^{(2)}_{14} S_1 S_4 - J^{(2)}_{23} S_2 S_3 - J^{(2)}_{24} S_2 S_4 - J^{(2)}_{34} S_3 S_4 + C$$

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)

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



Visible : **data**

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

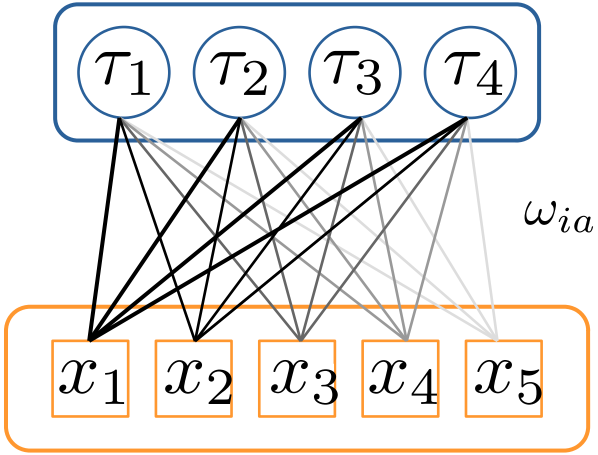
Universal approximator !

Le Roux and Bengio. Neural computation (2008)

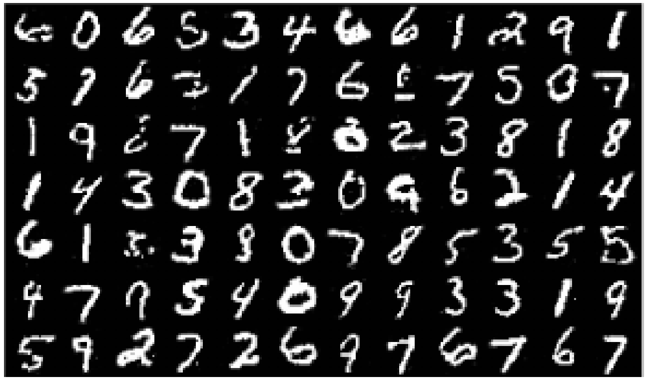
The Restricted Boltzmann Machine

-Smolensky, P. (1986)

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



B3 **Samples generated with the RBM**



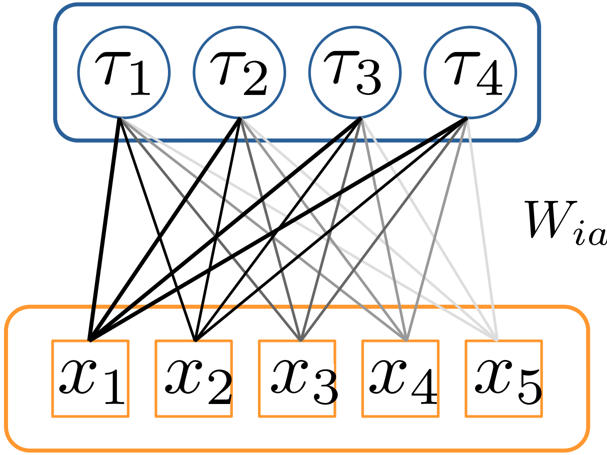
Universal approximator !

Le Roux and Bengio. Neural computation (2008)

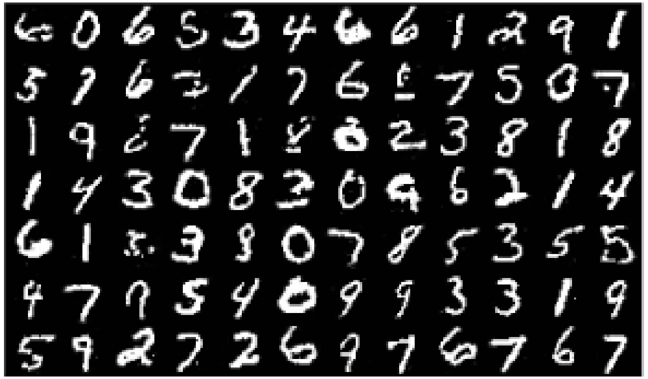
The Restricted Boltzmann Machine

-Smolensky, P. (1986)

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



B3 **Samples generated with the RBM**



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

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

$$\sigma_j, \tau_i \in \{\pm 1\}$$

Both Ising variables

$$\mathcal{H}_{RBM}(\boldsymbol{\sigma}) = -\log \sum_{\boldsymbol{\tau}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{\sigma}, \boldsymbol{\tau})} = -\sum_i \eta_i \sigma_i - \sum_a \log\cosh\left(\theta_a + \sum_i W_{ia} \sigma_i\right) + C$$

The RBM as a model for interacting spins

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

$\sigma_j, \tau_i \in \{\pm 1\}$

Both Ising variables

$$\mathcal{H}_{RBM}(\boldsymbol{\sigma}) = -\log \sum_{\boldsymbol{\tau}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{\sigma}, \boldsymbol{\tau})} = -\sum_i \eta_i \sigma_i - \sum_a \log\cosh\left(\theta_a + \sum_i W_{ia} \sigma_i\right) + C$$

$$= -\sum_j H_j \sigma_j - \sum_{j_1 > j_2} J^{(2)}_{j_1 j_2} \sigma_{j_1} \sigma_{j_2} - \sum_{j_1 > j_2 > j_3} J^{(3)}_{j_1 j_2 j_3} \sigma_{j_1} \sigma_{j_2} \sigma_{j_3} + \dots$$

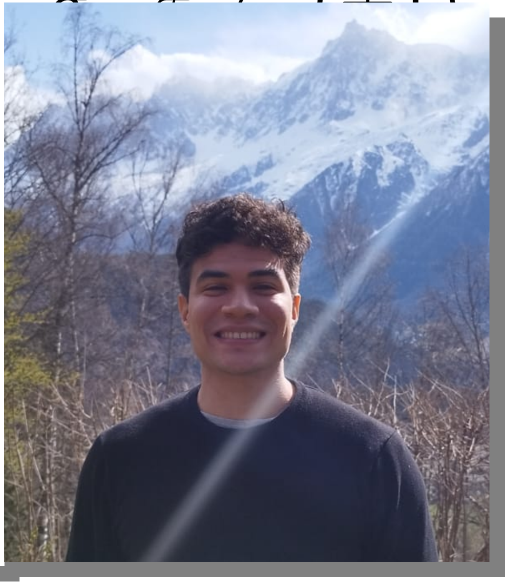
The RBM as a model for interacting spins

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

Sci|Post SciPost Phys. 16, 095 (2024)

Inferring effective couplings with restricted Boltzmann machines

Aurélien Decelle[1,2], Cyril Furtlehner[2],
Alfonso De Jesús Navas Gómez[1⋆] and Beatriz Seoane[2]



The RBM as a model for interacting spins

From the RBM to a generalized Ising model

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_j \eta_j \sigma_i - \sum_a \log\cosh\left(\sum_j w_{ja}\sigma_j + \zeta_a\right).$$

$$= -\sum_j \eta_j \sigma_j - \sum_{\boldsymbol{\sigma'}} \prod_j \delta_{\sigma_j \sigma'_j} \sum_a \ln\cosh\left(\sum_j w_{ja}\sigma'_j + \zeta_a\right).$$

$$= -\sum_j \eta_j \sigma_j - \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma'}} \prod_j (1 + \sigma_j \sigma'_j) \sum_a \ln\cosh\left(\sum_j w_{ja}\sigma'_j + \zeta_a\right).$$

From the RBM to a generalized Ising model

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_j \eta_j \sigma_i - \sum_a \log\cosh\left(\sum_j w_{ja}\sigma_j + \zeta_a\right).$$

$$= -\sum_j \eta_j \sigma_j - \sum_{\boldsymbol{\sigma}'} \prod_j \delta_{\sigma_j \sigma'_j} \sum_a \ln\cosh\left(\sum_j w_{ja}\sigma'_j + \zeta_a\right).$$

$$= -\sum_j \eta_j \sigma_j - \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \prod_j (1 + \sigma_j \sigma'_j) \sum_a \ln\cosh\left(\sum_j w_{ja}\sigma'_j + \zeta_a\right).$$

$$(1+\sigma_1\sigma'_1)(1+\sigma_2\sigma'_2)\cdots(1+\sigma_{N_v}\sigma'_{N_v}) = 1 + \sum_j \sigma_j\sigma'_j + \sigma_1\sigma_2\sigma'_1\sigma'_2 + \cdots + \sigma_1\sigma_2\sigma_3\sigma'_1\sigma'_2\sigma'_3 + \cdots$$

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

From the RBM to a generalized Ising model

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

$$H_j = \eta_j + \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \sum_i \sigma'_j \ln\cosh\left(\sum_k w_{ik}\sigma'_k + \zeta_i\right)$$

$$J^{(n)}_{j_1\ldots j_n} = \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \sum_i \sigma'_{j_1}\ldots\sigma'_{j_n} \ln\cosh\left(\sum_k w_{ik}\sigma'_k + \zeta_i\right)$$

$$= -\sum_j \eta_j\sigma_j - \frac{1}{2^{N_\mathrm{v}}} \sum_{\boldsymbol{\sigma}'} \prod_j (1+\sigma_j\sigma'_j) \sum_a \ln\cosh\left(\sum_j w_{ja}\sigma'_j + \zeta_a\right).$$

$$(1+\sigma_1\sigma'_1)(1+\sigma_2\sigma'_2)\cdots(1+\sigma_{N_v}\sigma'_{N_v}) = 1 + \sum_j \sigma_j\sigma'_j + \sigma_1\sigma_2\sigma'_1\sigma'_2 + \cdots + \sigma_1\sigma_2\sigma_3\sigma'_1\sigma'_2\sigma'_3 + \cdots$$

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

From the RBM to a generalized Ising model

Introduce the random variable

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

N_v large

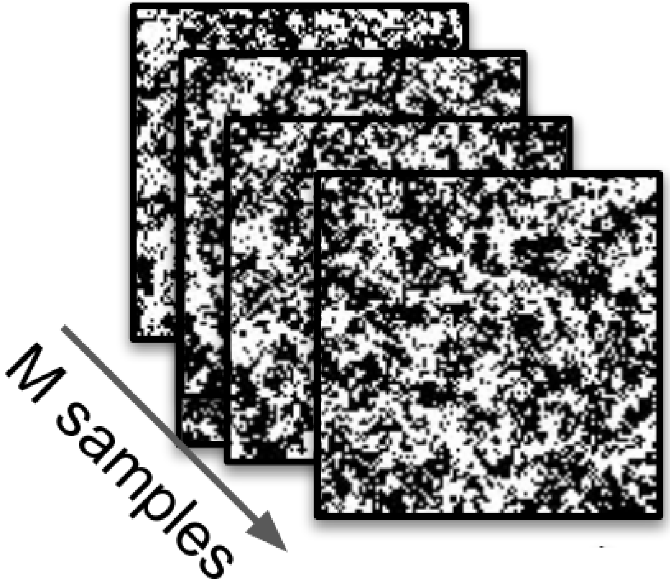
Central limit theorem

$$H_j = \eta_j + \frac{1}{2} \sum_a \mathbb{E}_{X_a^{(j)}} \left[\ln \frac{\cosh\left(\zeta_a + w_{ja} + X_a^{(j)}\right)}{\cosh\left(\zeta_a - w_{ja} + X_a^{(j)}\right)} \right]$$

$$J_{j_1 j_2}^{(2)} = \frac{1}{4} \sum_a \mathbb{E}_{X_a^{(j_1 j_2)}} \left[\ln \frac{\cosh\left(\zeta_a + w_{j_1 a} + w_{j_2 a} + X_i^{(j_1 j_2)}\right) \times \cosh\left(\zeta_a - (w_{j_1 a} + w_{j_2 a}) + X_a^{(j_1 j_2)}\right)}{\cosh\left(\zeta_a + (w_{j_1 a} - w_{j_2 a}) + X_a^{(j_1 j_2)}\right) \times \cosh\left(\zeta_a - (w_{a j_1} - w_{a j_2}) + X_a^{(j_1 j_2)}\right)} \right]$$

Numerical controlled experiments

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

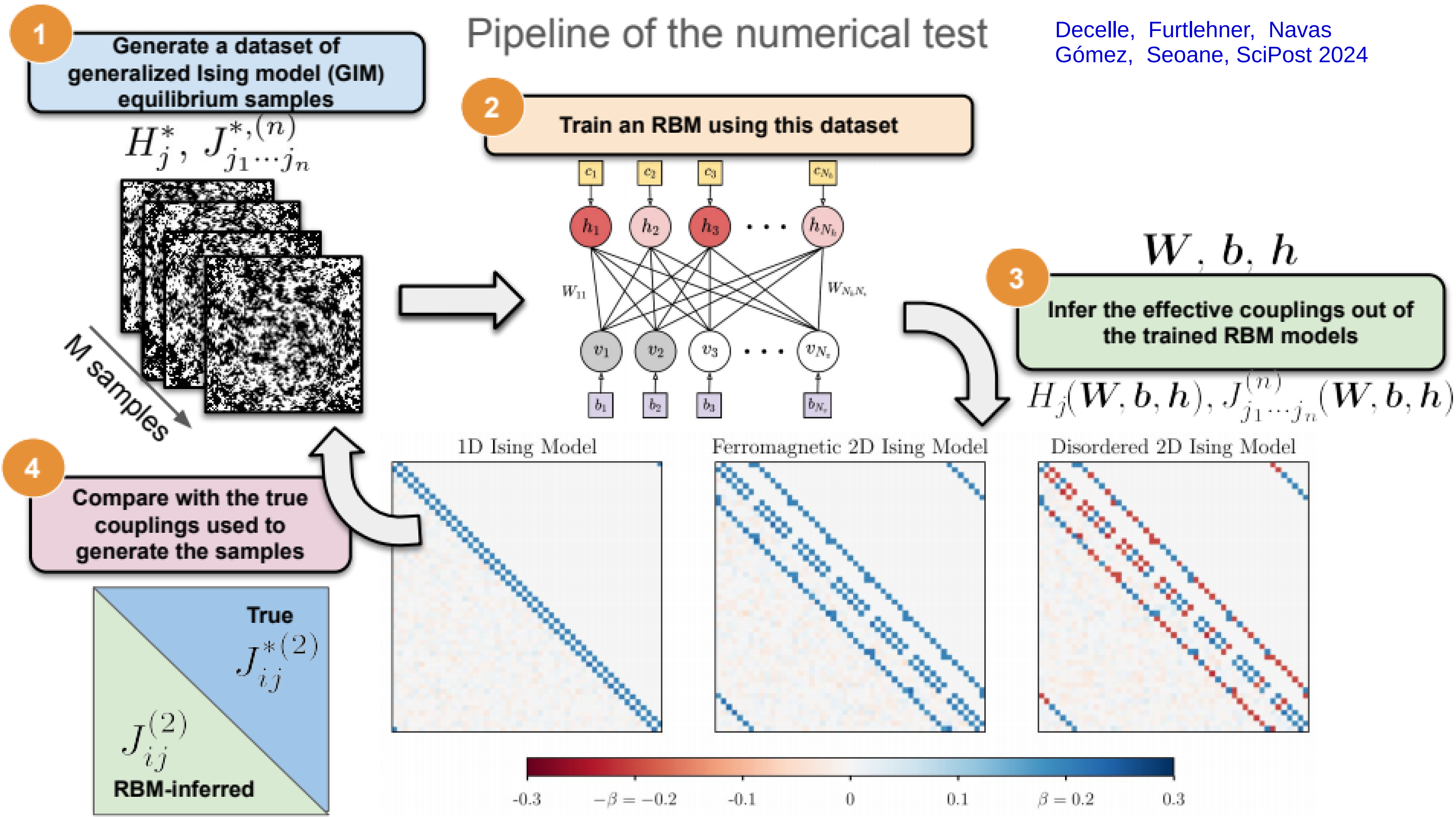


$$\beta = \frac{1}{T}$$

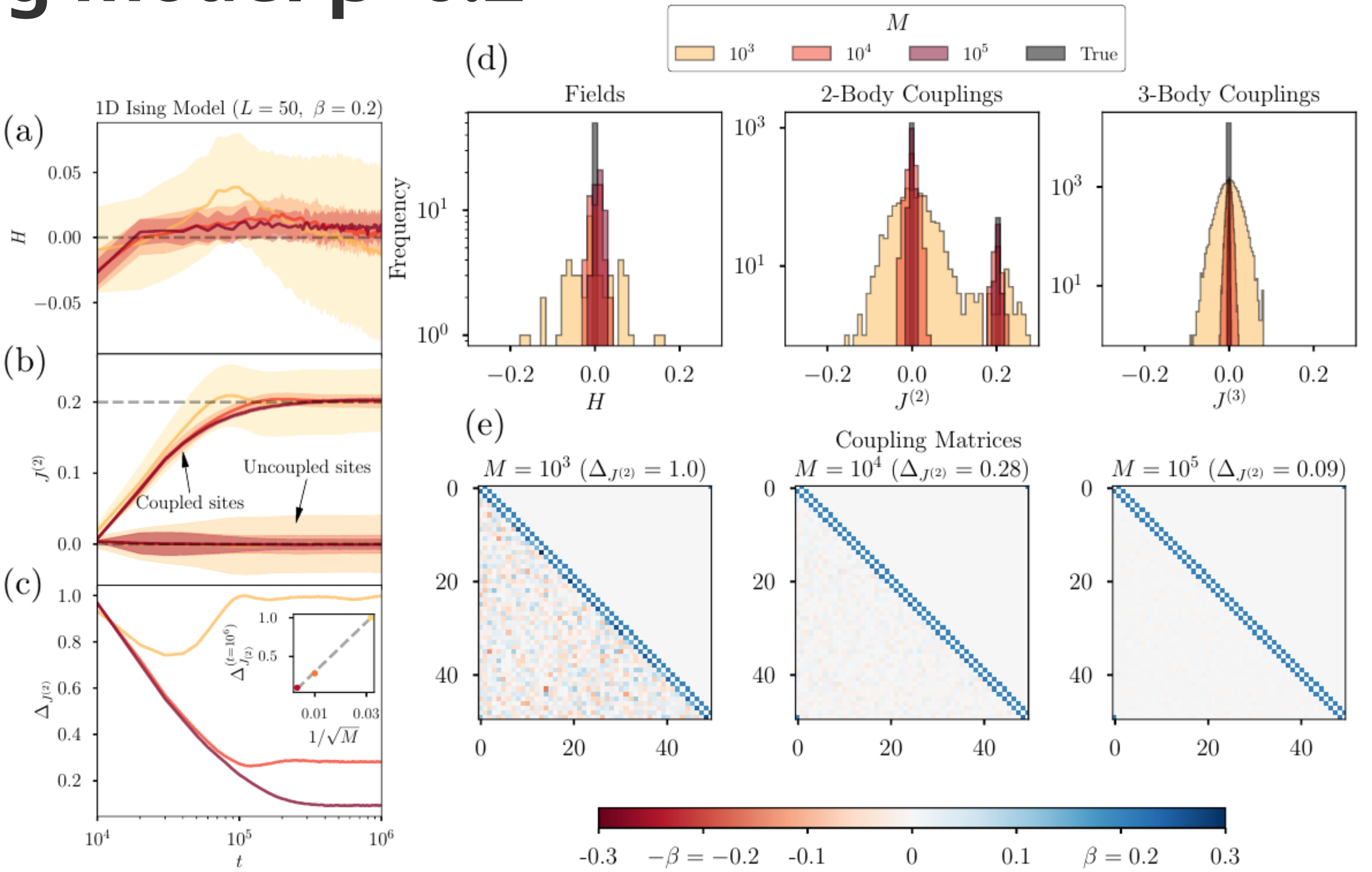
Generate equilibrium samples
With a known model

Pipeline of the numerical test

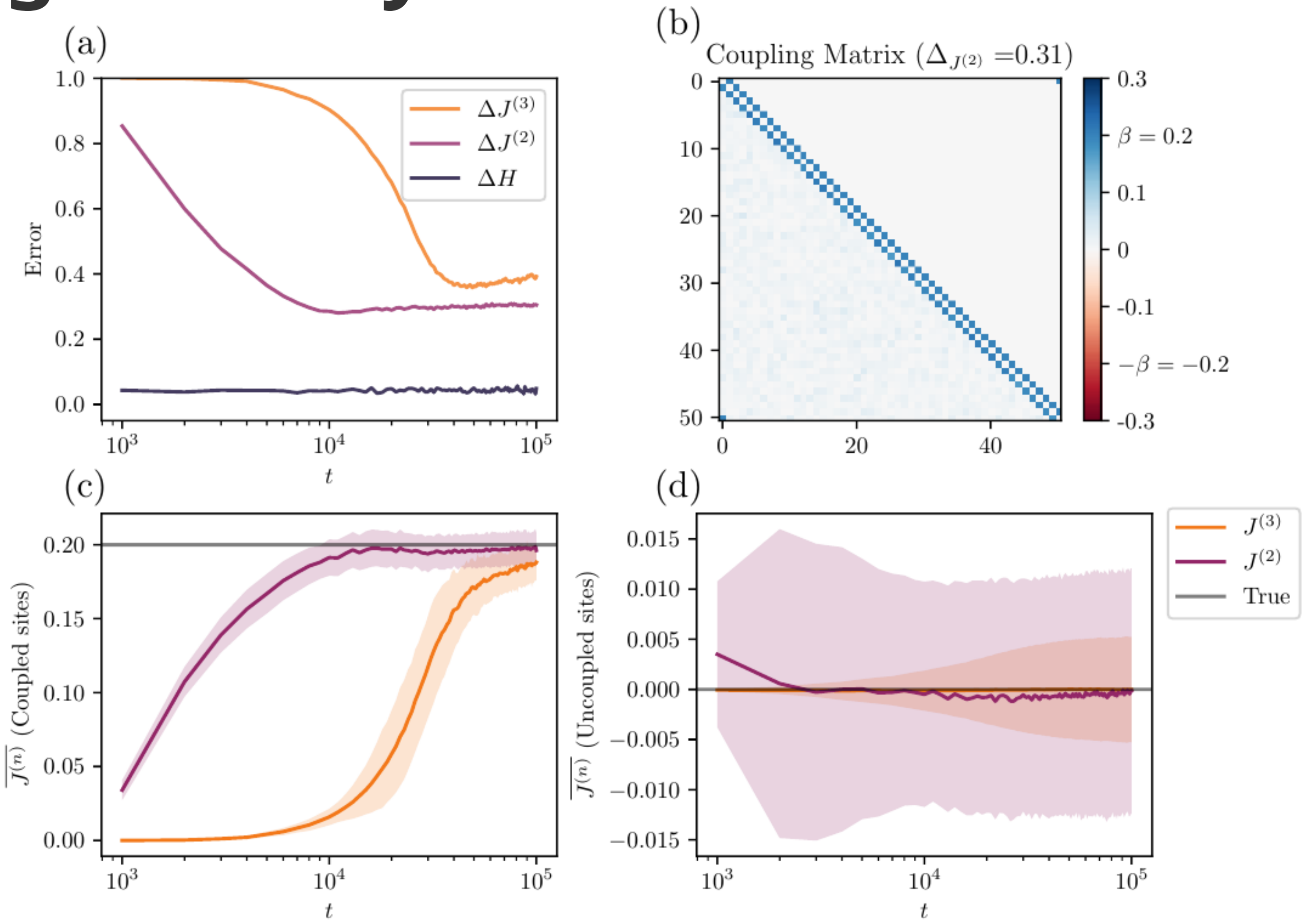
Decelle, Furtlehner, Navas Gómez, Seoane, SciPost 2024



1D Ising model β=0.2

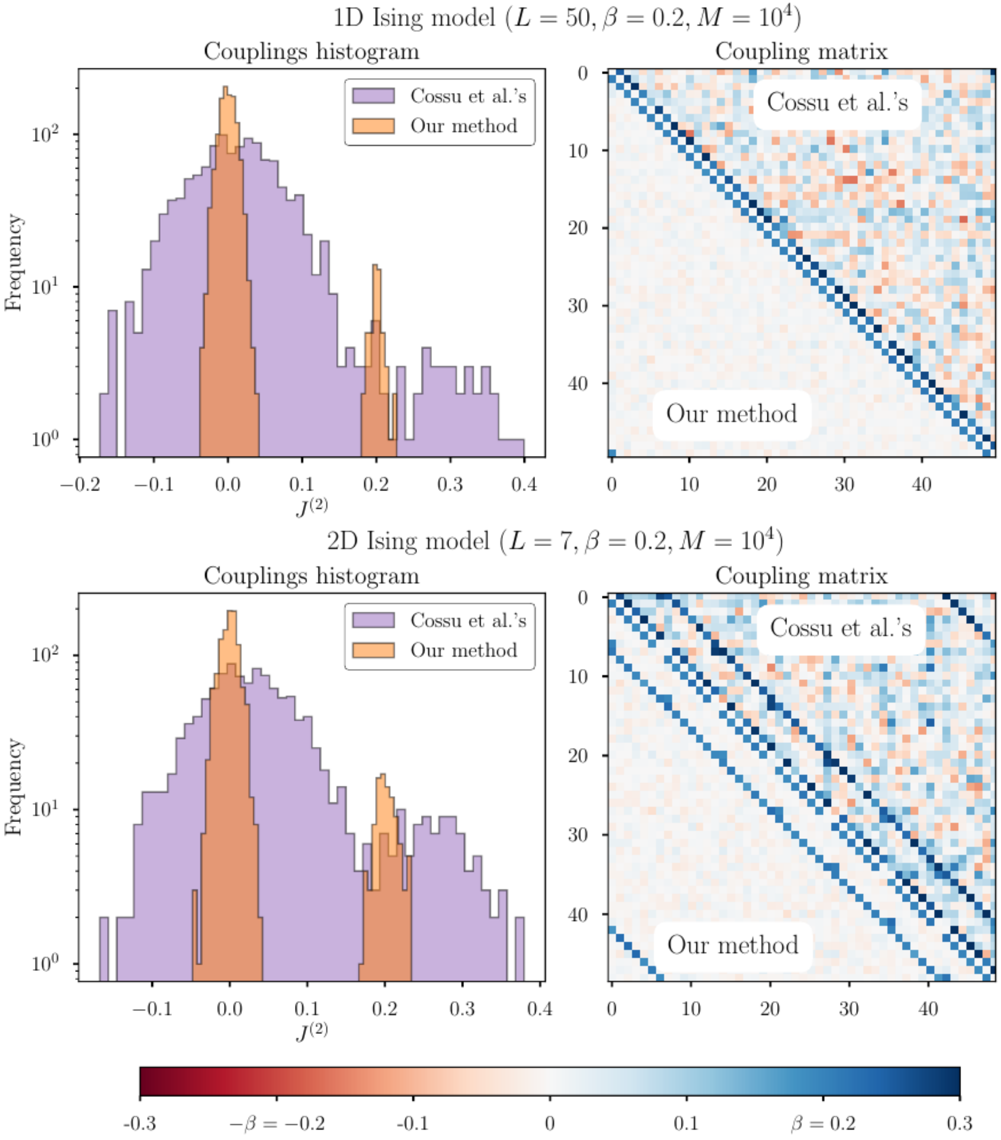


1D Ising + 3-body interactions



Previous attempts

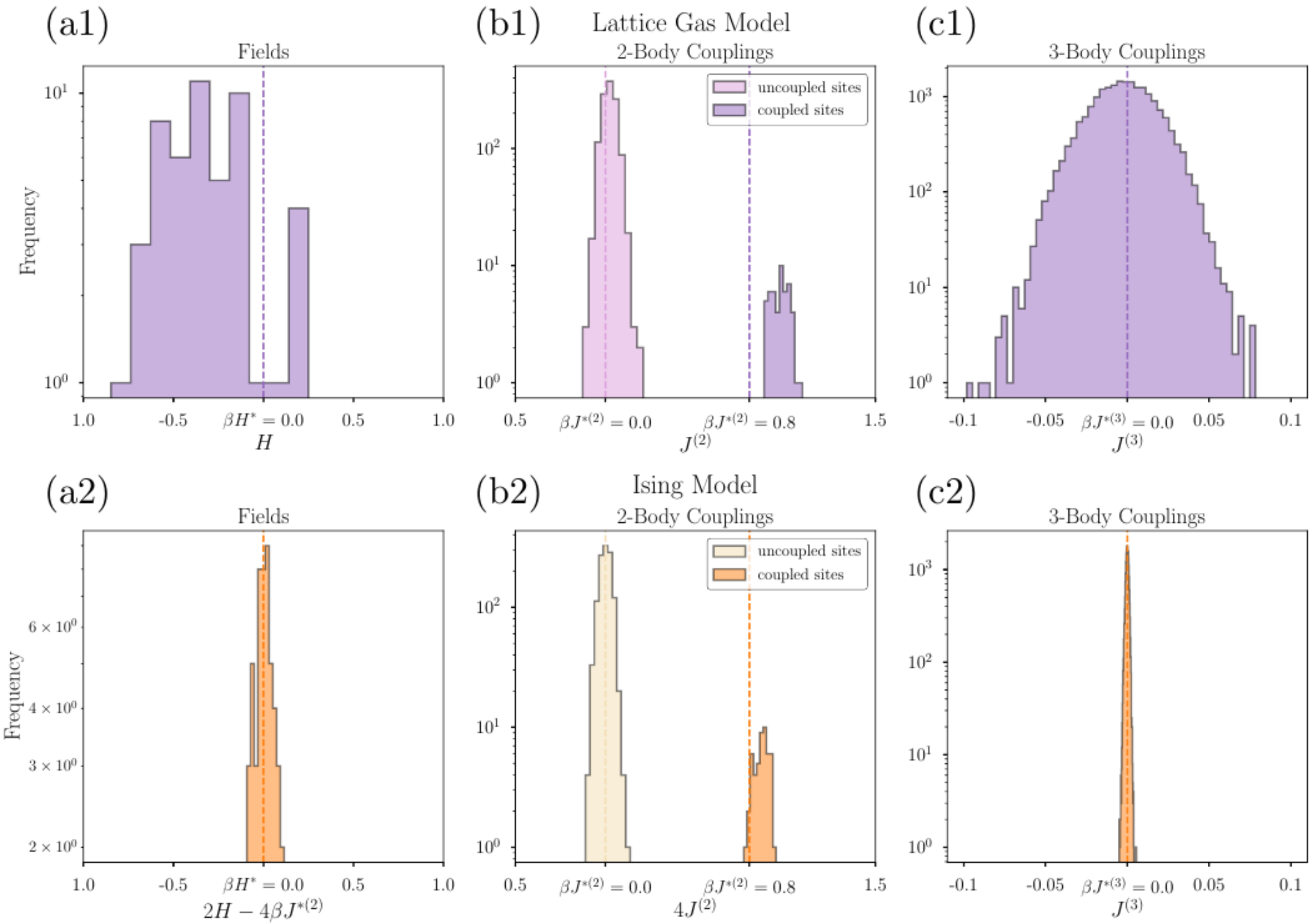
G. Cossu, L. Del Debbio, T. Giani, A. Khamseh and M. Wilson, Phys. Rev. B (2019)



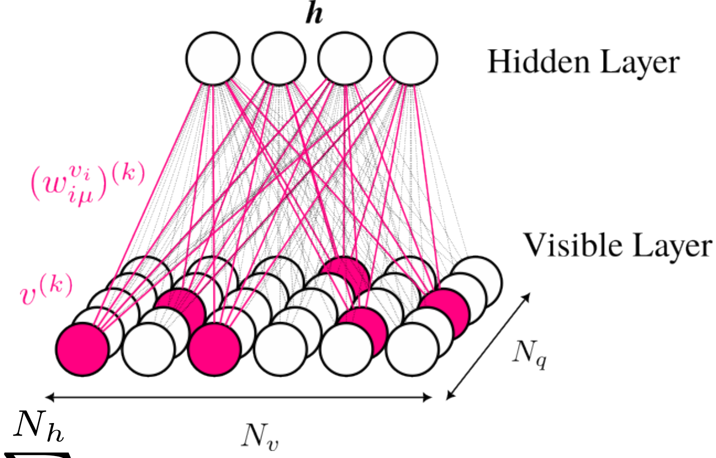
Previous attempts

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

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



Beyond Ising spins



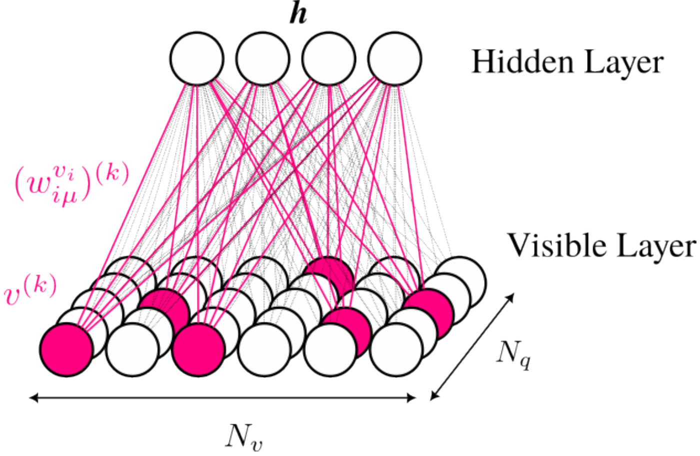
One can generalize to Potts variables

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

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

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

From Ising to Potts



We can use it to infer

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

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

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

Main difficulty: gauge symmetry

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

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 !

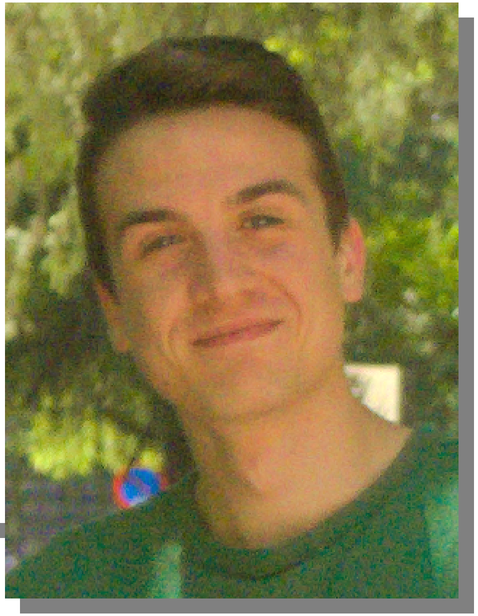
And the zero sum gauge in the RBM is not equivalent to the zero sum gauge in the effective Potts model

PHYSICAL REVIEW E **108**, 014110 (2023)

Unsupervised hierarchical clustering using the learning dynamics of restricted Boltzmann machines

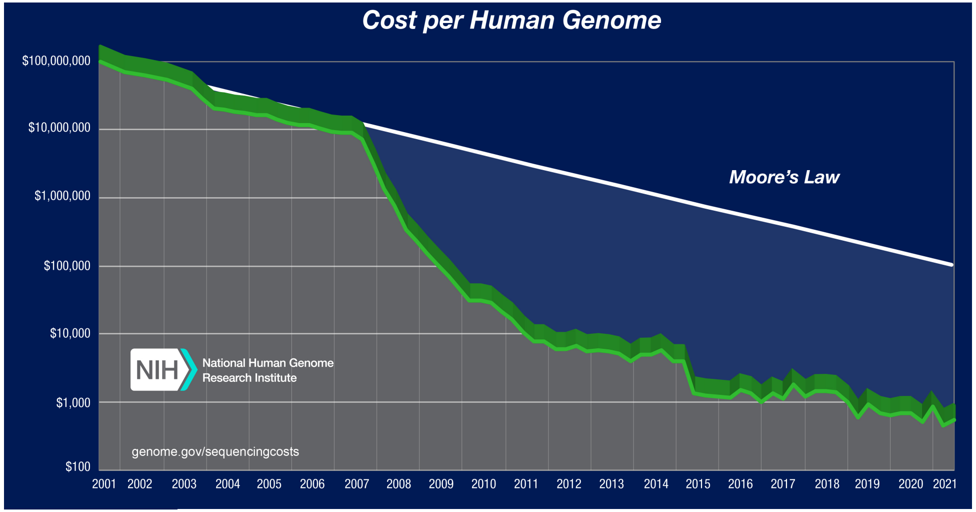
Aurélien Decelle and Beatriz Seoane
Departamento de Física Teórica, Universidad Complutense de Madrid, 28040 Madrid, Spain and Université Paris-Saclay, CNRS, INRIA Tau team, LISN, 91190 Gif-sur-Yvette, France

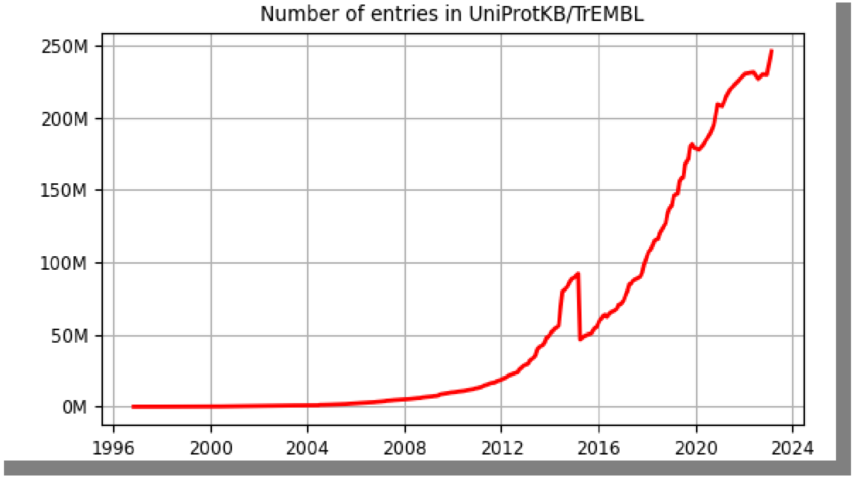
Lorenzo Rosset*
Departamento de Física Teórica, Universidad Complutense de Madrid, 28040 Madrid, Spain

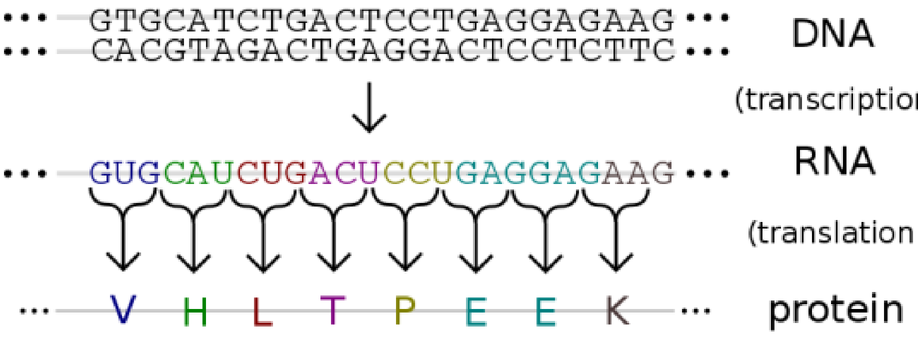


Analyzing the free energy landscape

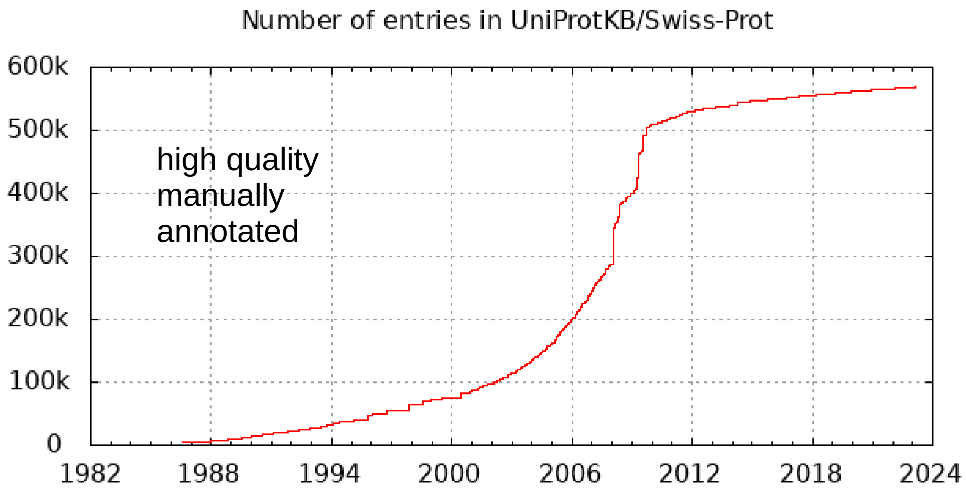
Motivation

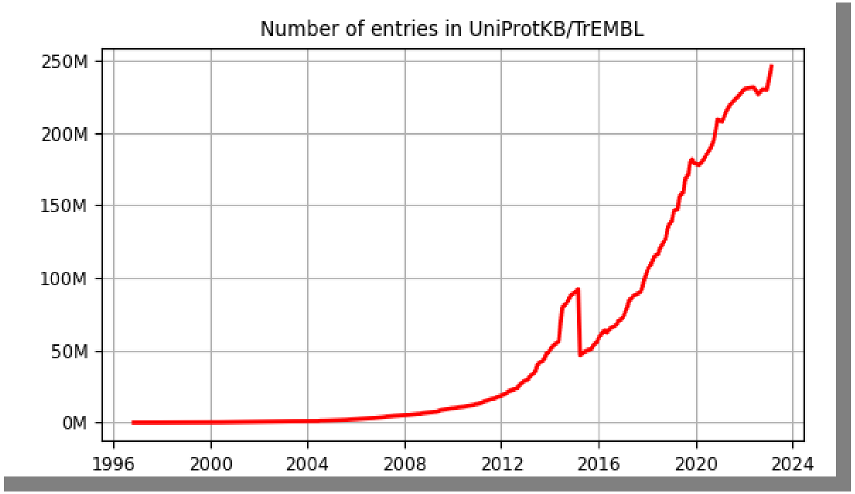

Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

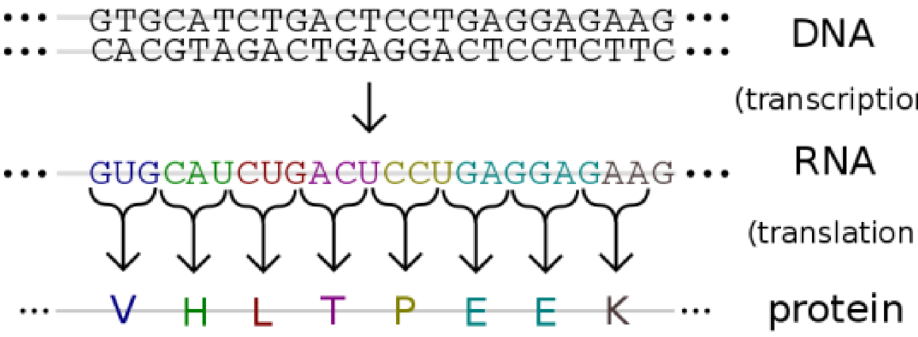

Number of entries in UniProtKB/TrEMBL
250M
200M
150M
100M
50M
0M
1996 2000 2004 2008 2012 2016 2020 2024


GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC
DNA
(transcriptio
GUGCAUCUGACUCCUGAGGAGAAG
RNA
(translation
V H L T P E E K
protein

Motivation

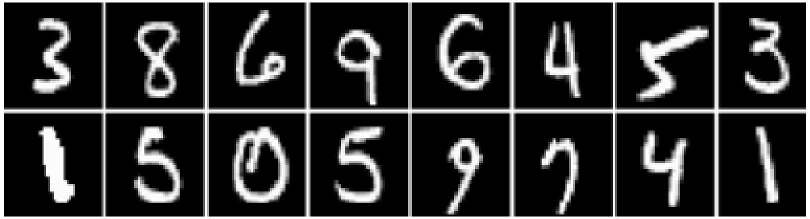

Number of entries in UniProtKB/Swiss-Prot
high quality
manually
annotated
600k
500k
400k
300k
200k
100k
0
1982
1988
1994
2000
2006
2012
2018
2024


Number of entries in UniProtKB/TrEMBL
250M
200M
150M
100M
50M
0M
1996
2000
2004
2008
2012
2016
2020
2024


GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC
DNA
(transcriptio
GUGCAUCUGACUCCUGAGGAGAAG
RNA
(translation
V H L T P E E K
protein

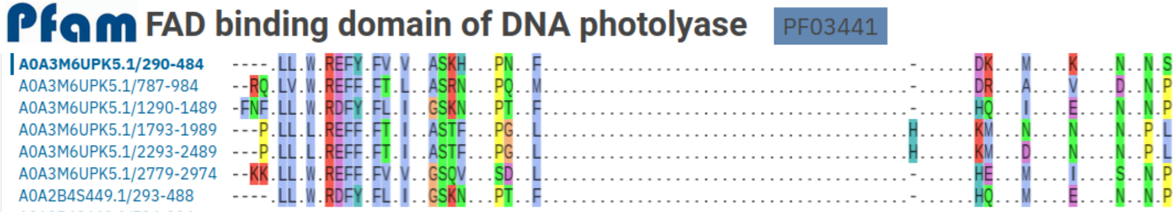
We need tools to automatically tag data

MNIST



digit →

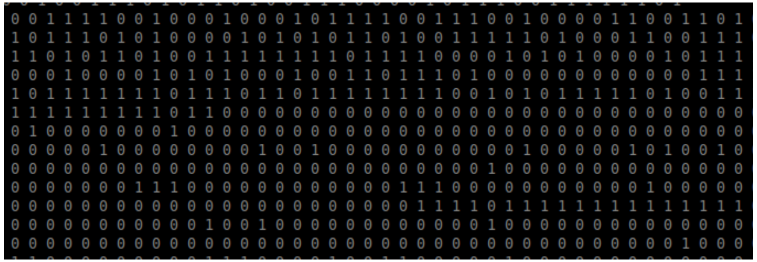
3, 8, 6, 9, 6, 4, 5, 3,
1, 5, 0, 5, 9, 7, 4, 1



Biological function →

Human Genome dataset → mutations genome

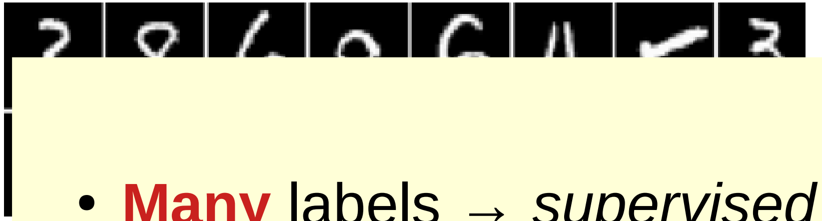
A global reference for human genetic variation, Nature 526(7571),68 (2015),



Population origin →

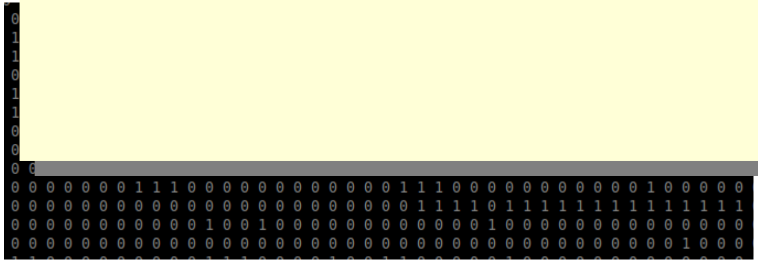
We need tools to automatically tag data

MNIST

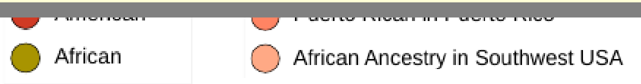


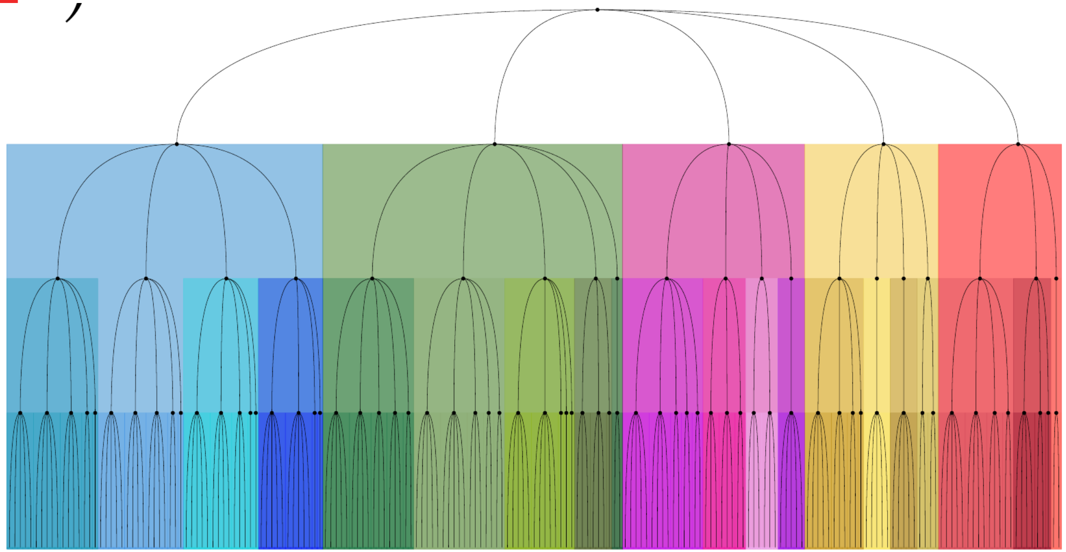
- **Many** labels → *supervised* learning

- **None** or **so few** labels → *unsupervised* or (*semi supervised*) learning



origin



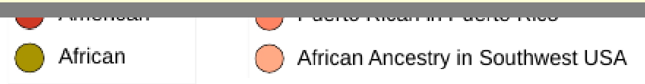


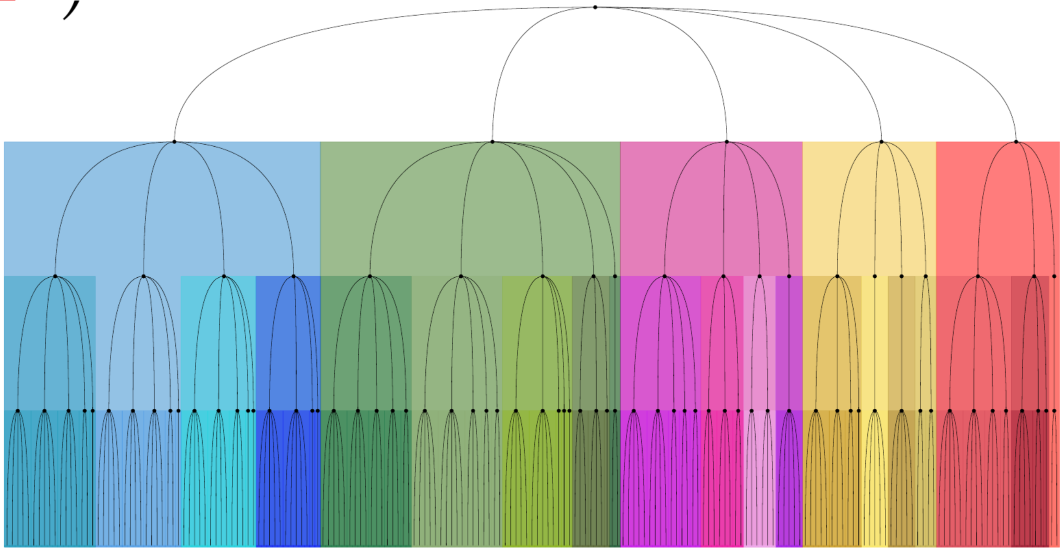
Evolutionary process

- **<u>None</u>** or **<u>so few</u>** <u>labels</u> → *unsupervised* or (*semi supervised*) learning

Detect families and subfamilies in the data → **<u>hierarchical</u>** ***clustering***

- **<u>Curse of dimensionality</u>**



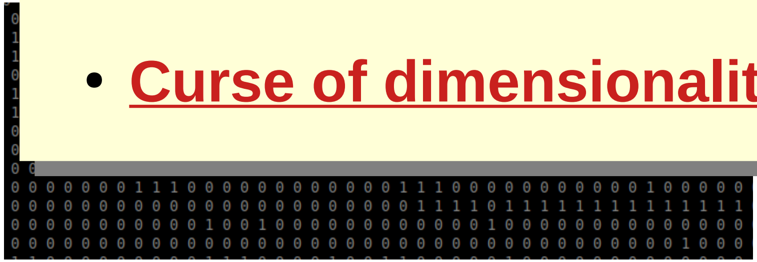


$$p_{\mathcal{D}}(\boldsymbol{x}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

- **None** or **so few** labels → *unsupervised* or (*semi supervised*) learning

Detect families and subfamilies in the data → **hierarchical** ***clustering***

- **Curse of dimensionality**



Step 0 : Principal Component Analysis

Human Genome dataset → mutations genome

A global reference for human genetic variation, Nature 526(7571),68 (2015),

Population origin ?

Mutation sites

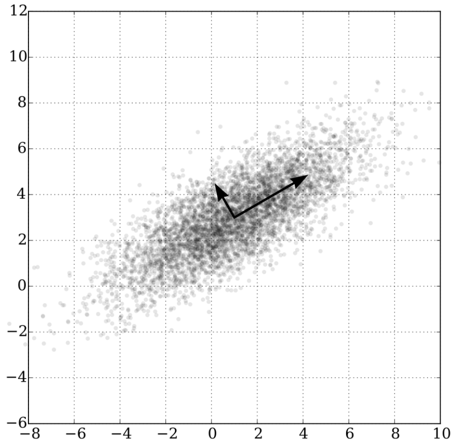
N_v

Human individuals

M

$$\begin{matrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \boldsymbol{X}^{(1)} \\ 0 & 1 & 0 & 1 & \dots & 0 & 1 & 0 & 0 & \boldsymbol{X}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & \dots & 0 & 1 & 0 & 1 & \boldsymbol{X}^{(M)} \end{matrix}$$

$$\sum = Cov[X_i, X_j]$$



Eigenvectors : $\boldsymbol{v}_\alpha$

Directions of maximal variation

Step 0 : Principal Component Analysis

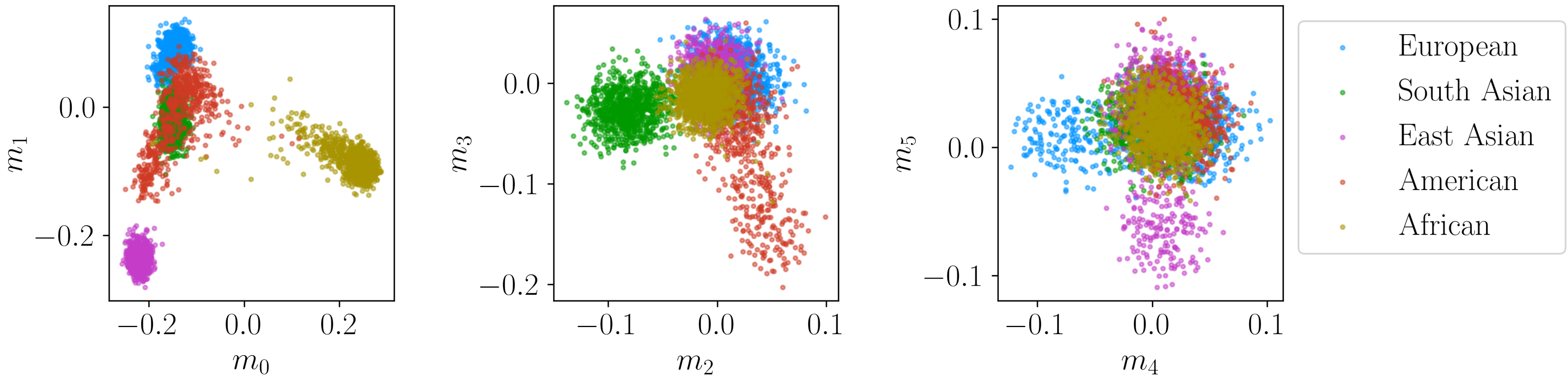
Human Genome dataset → mutations genome

A global reference for human genetic variation, Nature 526(7571),68 (2015),

Population origin ?

$$m_\alpha^{(i)} = \boldsymbol{v}_\alpha \cdot \boldsymbol{X}^{(i)}$$

PCA Human Genome

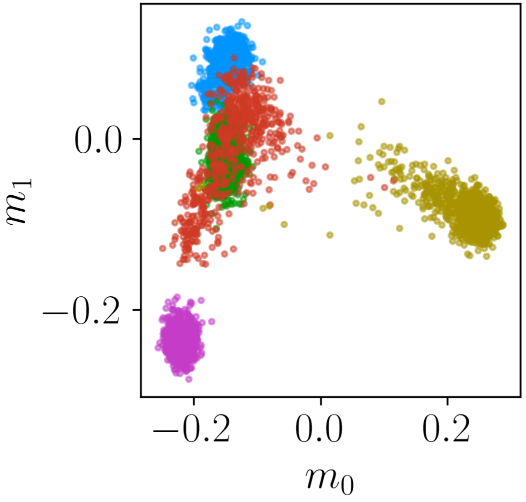


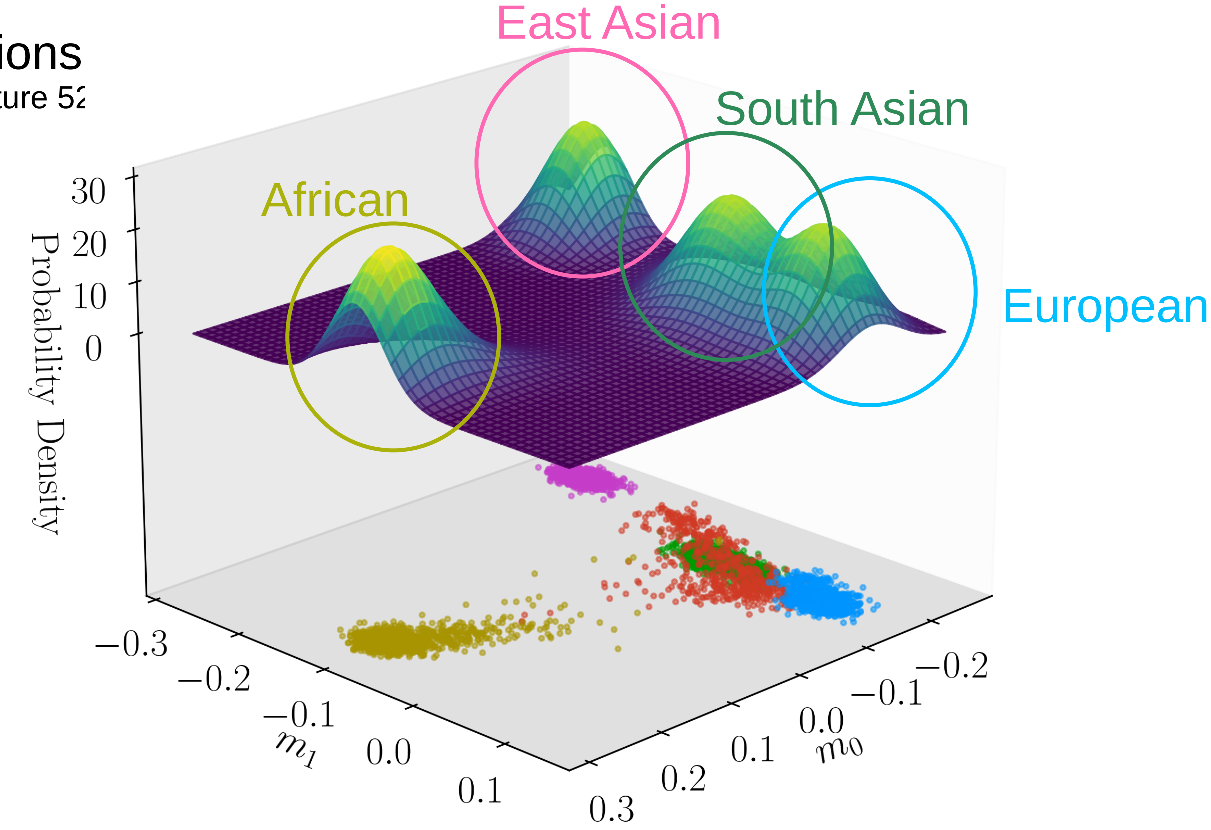
Step 0 : Principal Component Analysis

Human Genome dataset → mutations

A global reference for human genetic variation, Nature 52

$$m_\alpha^{(i)} = \boldsymbol{v}_\alpha \cdot \boldsymbol{X}^{(i)}$$



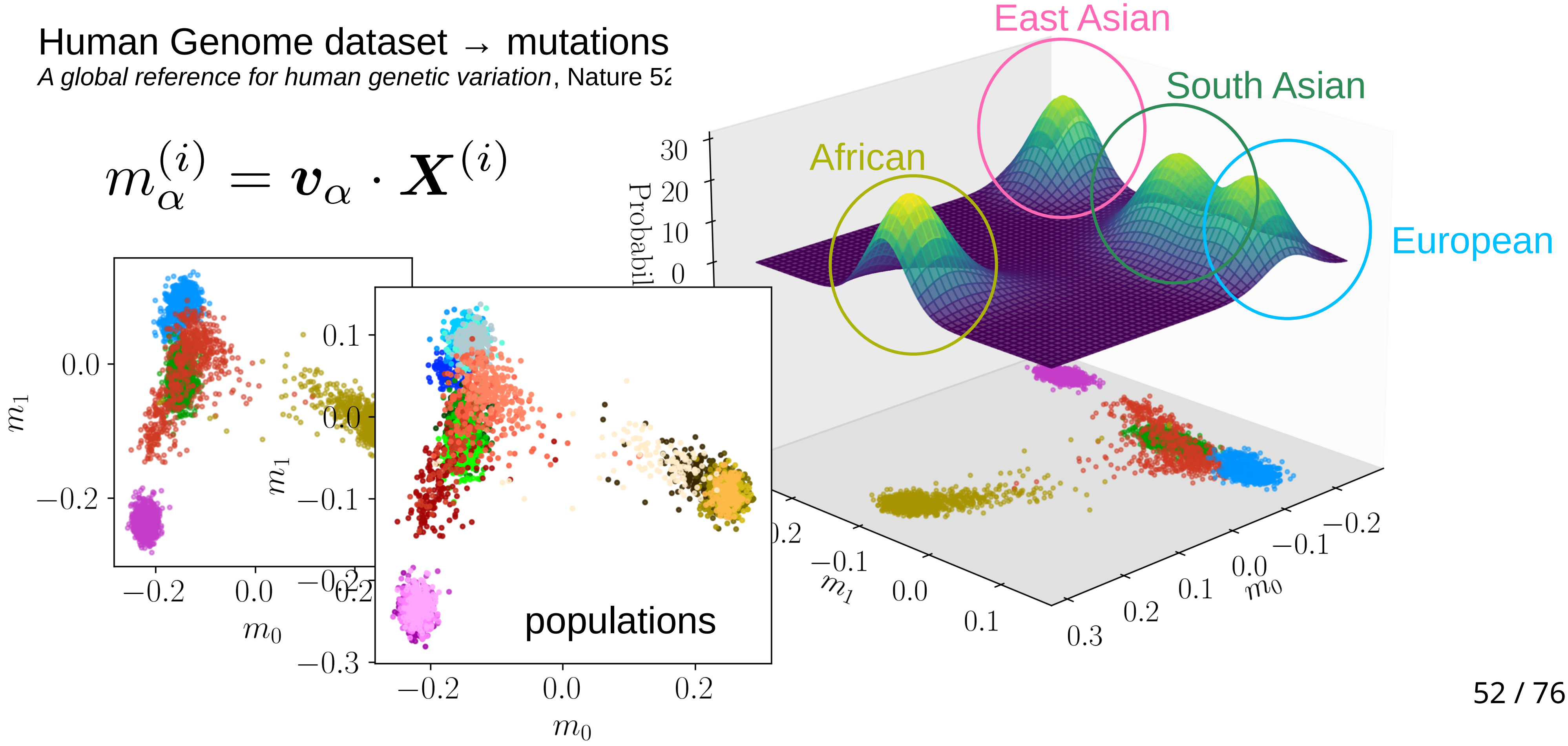


Step 0 : Principal Component Analysis

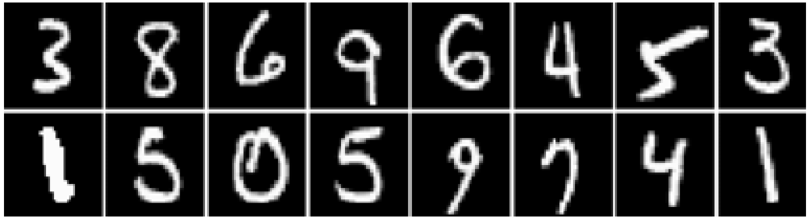
Human Genome dataset → mutations

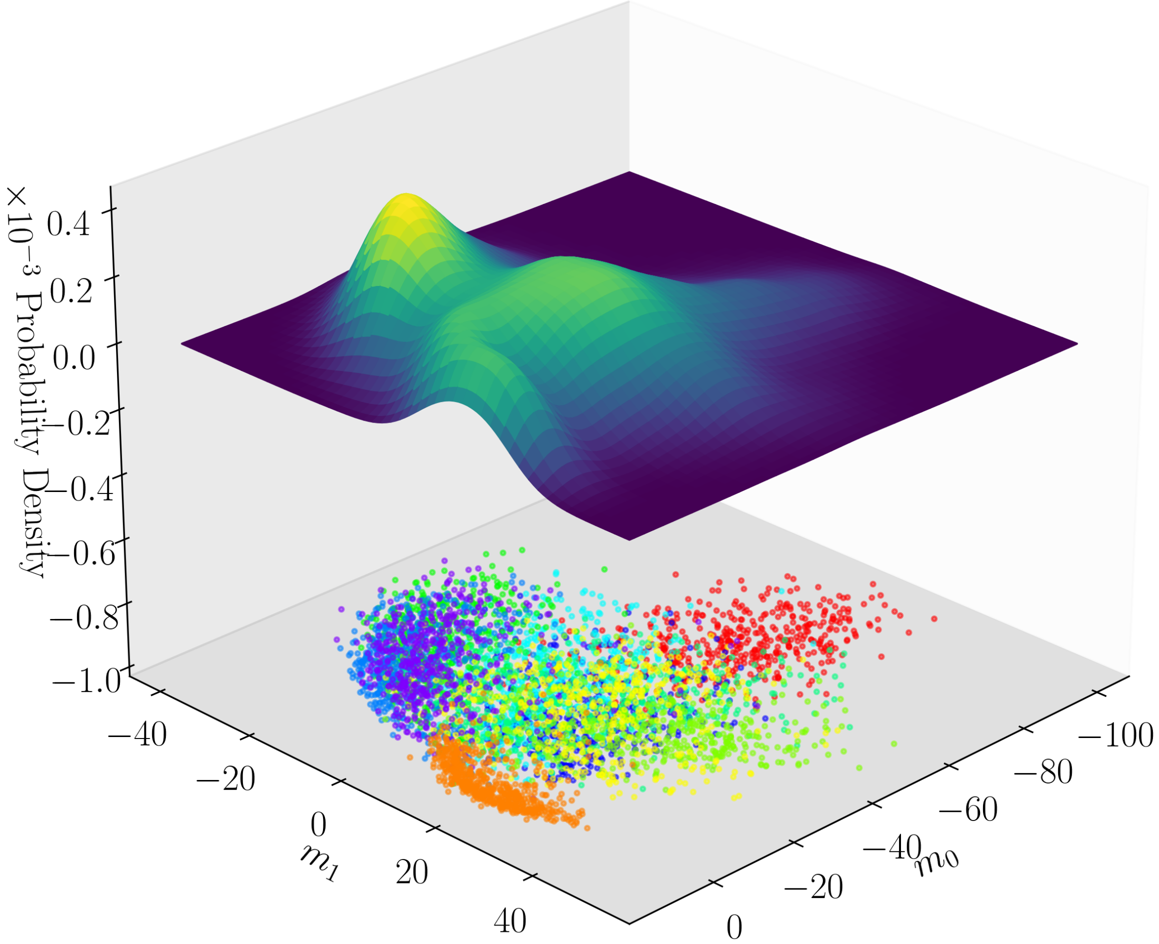
A global reference for human genetic variation, Nature 52

$$m_\alpha^{(i)} = \boldsymbol{v}_\alpha \cdot \boldsymbol{X}^{(i)}$$

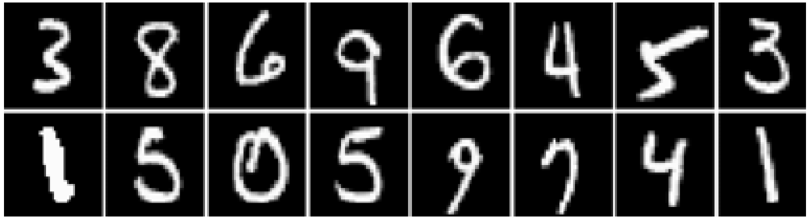


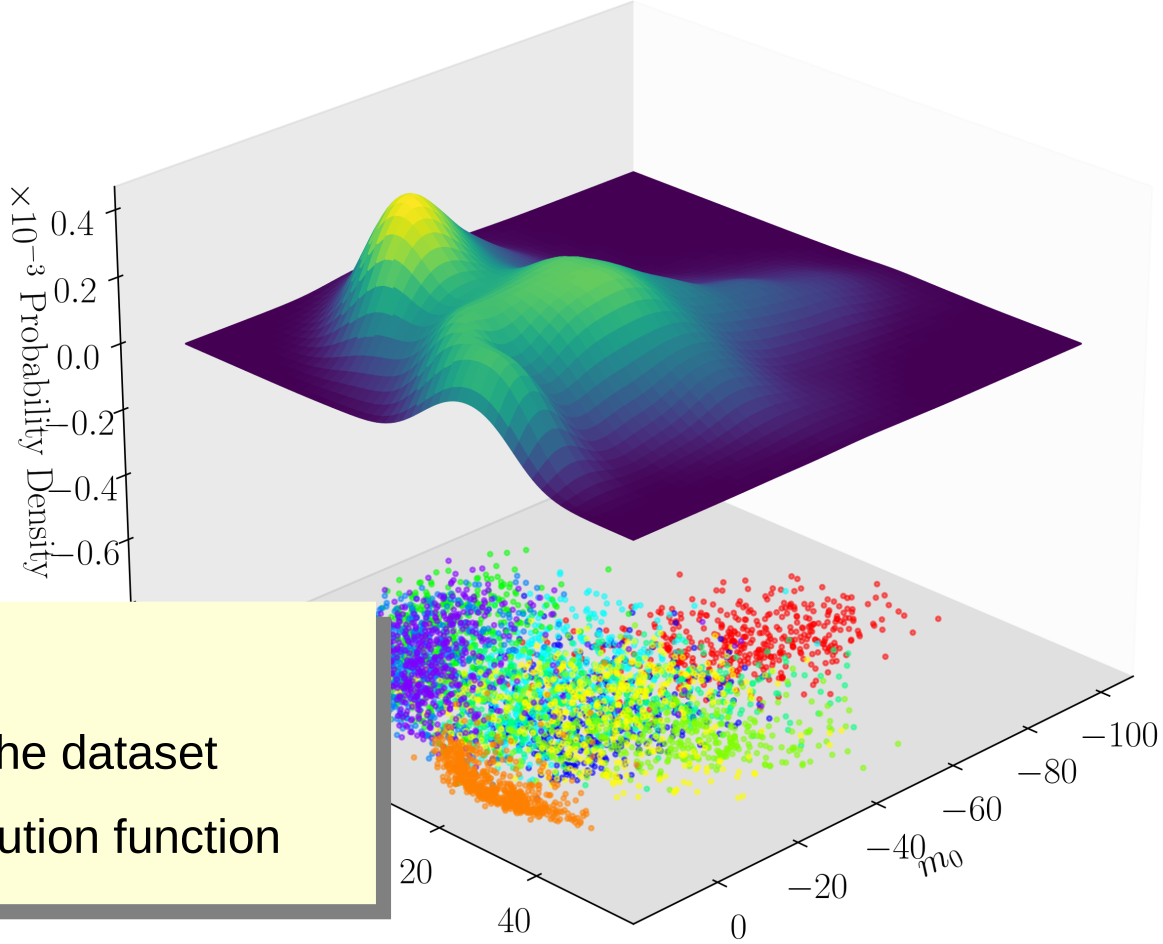
Step 0 : Principal Component Analysis





Step 0 : Principal Component Analysis





We need :

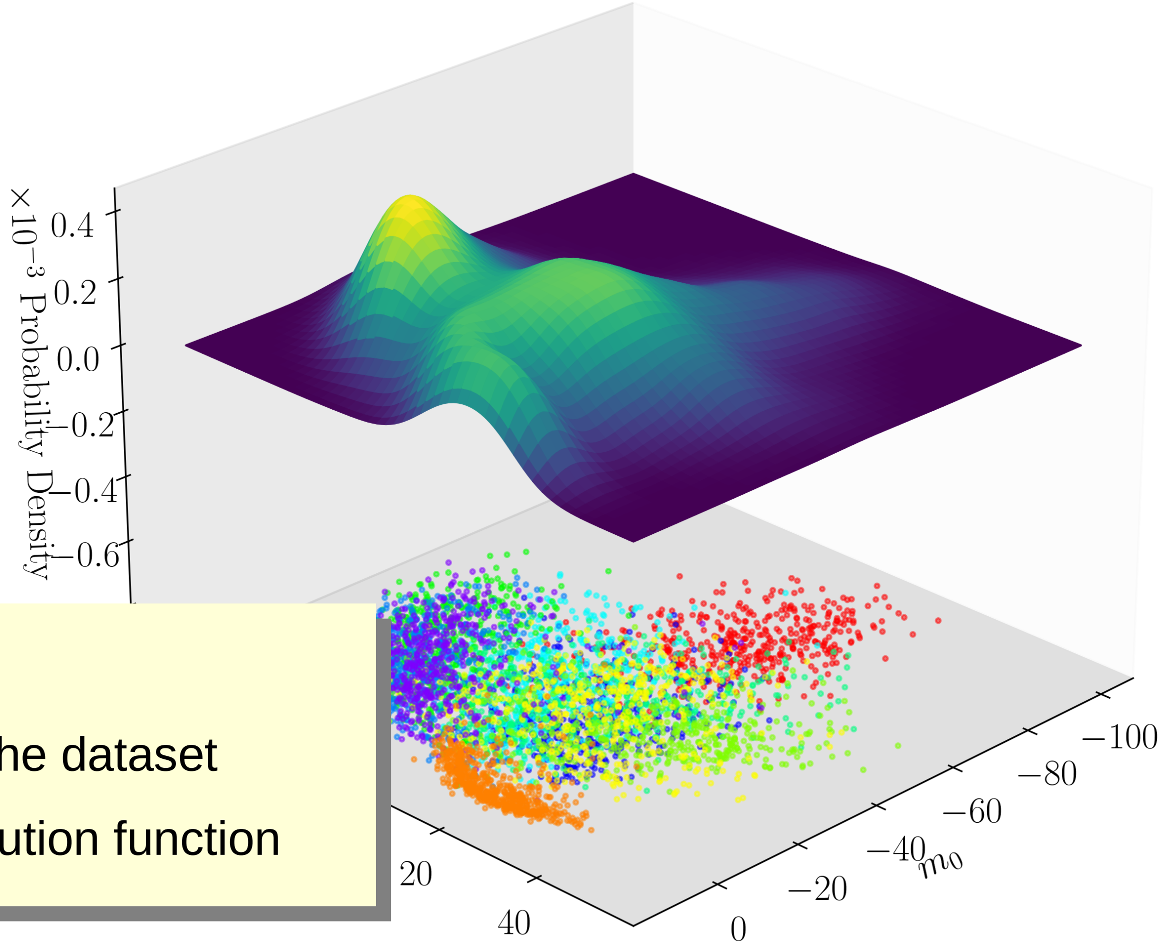
- Better decomposition (features) of the dataset
- Finer probe of the probability distribution function

Step 0 : Principal Component Analysis

We have a model for the probability

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

Can we probe the maxima?

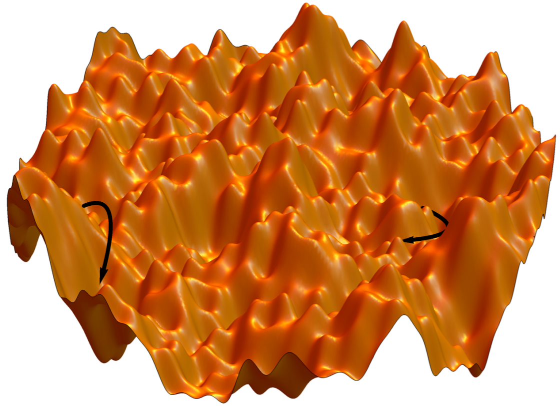


We need :

- Better decomposition (features) of the dataset
- Finer probe of the probability distribution function

Free energy landscape

$$p(\boldsymbol{S}) = \frac{e^{-E_{RBM}(\boldsymbol{S})}}{Z}$$



q^N Number of states but so few contribute

$$Z = \sum_{\{\boldsymbol{S}\}} e^{-E_{RBM}(\boldsymbol{S})} = \sum_U g(U) e^{-U} = \sum_U e^{S(U)-U} = \sum_U e^{-F(U)} = \sum_U e^{-Nf(U)}$$

$F = U - TS$ “Free energy”

The states with lower f(U) are those that dominate the measure

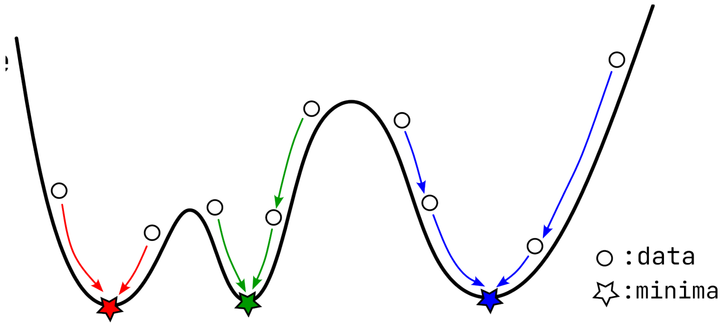
Free energy landscape

- 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 $\boldsymbol{M}$s with lower $f(\boldsymbol{M})$

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



[T. Plefka, J. Phys. A 15, 1971 (1982),
A. Georges and J. S. Yedidia, J. Phys. A 24, 2173 (1991)]

Approximate the free energy

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

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

[T. Plefka, J. Phys. A 15, 1971 (1982),
A. Georges and J. S. Yedidia, J. Phys. A 24, 2173 (1991)]

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

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

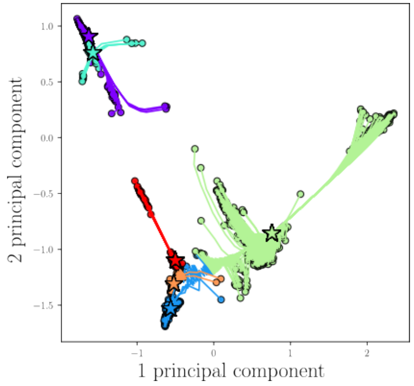
$$f_i^q[t+1] \leftarrow \mathrm{softmax}_q \left[a_i^q + \sum m_\mu[t+1] w_{i\mu}^q + \sum (m_\mu[t+1] - m_\mu^2[t+1]) \left(\frac{1}{2} (w_{i\mu}^q)^2 - w_{i\mu}^q \sum_p f_i^p[t] w_{i\mu}^p \right) \right]$$

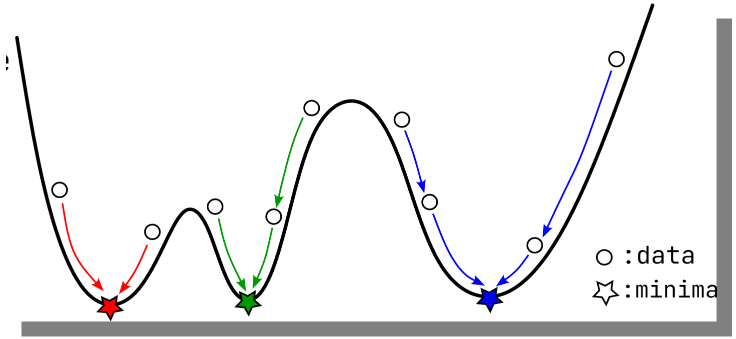
Solve iteratively

[Thouless, Anderson, Palmer, Philos. Mag. 35, 593 (1977)]

Approximate the free energy

-
-



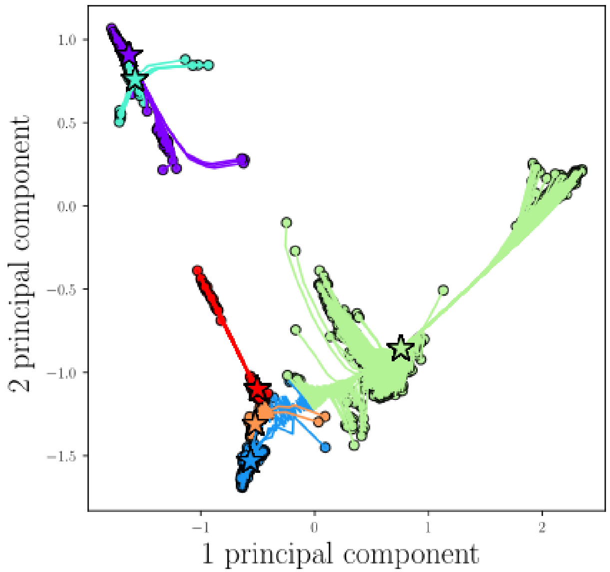


- Minima $\nabla f(M) = \mathbf{0}$ ⇒ set of self-consistent equations (TAP eqs.)

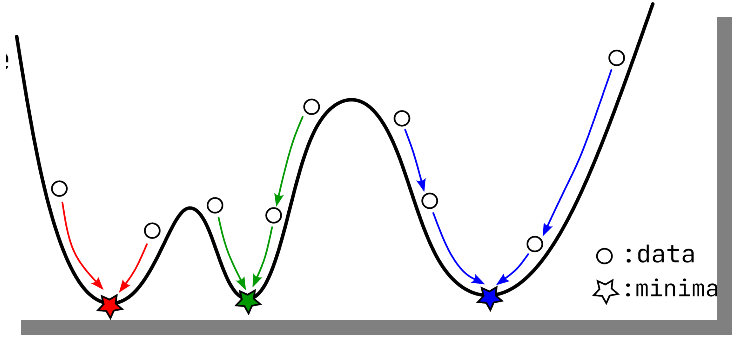
$$m_\mu[t+1] \leftarrow \mathrm{sigmoid}\left[b_\mu + \sum_{iq} f_i^q[t] w_{i\mu}^q + \left(m_\mu[t] - \frac{1}{2} \right) \left(\sum_i \left(\sum_q f_i^q[t] w_{i\mu}^q \right)^2 - \sum_{iq} (w_{i\mu}^q)^2 f_i^q[t] \right) \right]$$

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

Solve iteratively



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



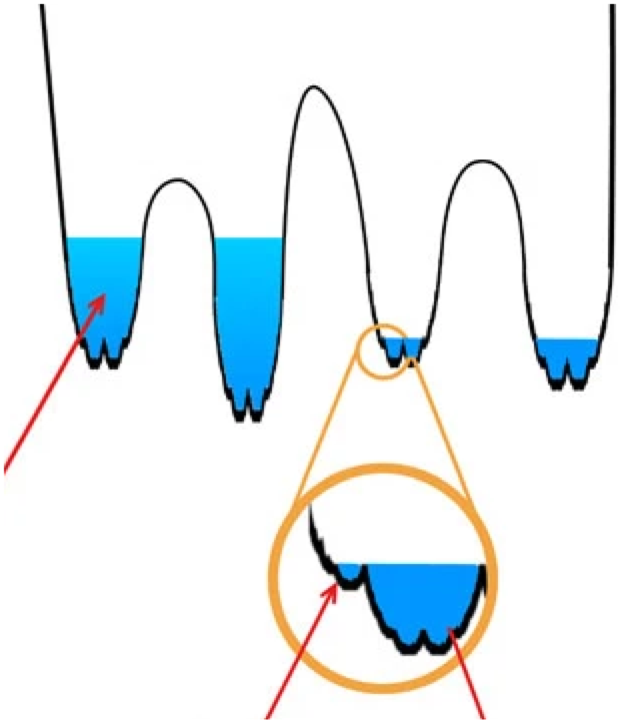
- Minima $\nabla f(\boldsymbol{M}) = \boldsymbol{0}$ ⇒ set of self-consistent equations (TAP eqs.)

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

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

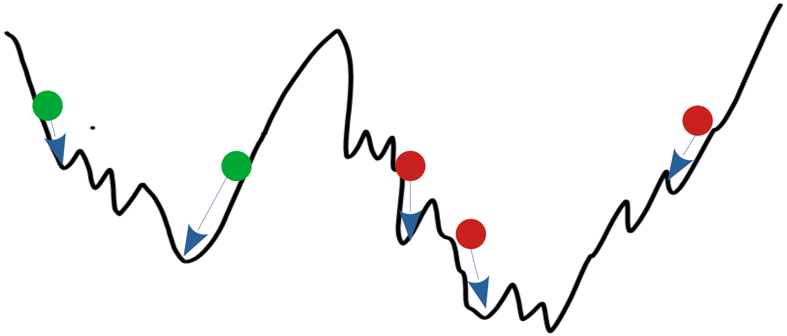
Solve iteratively

Data has a hierarchical organization

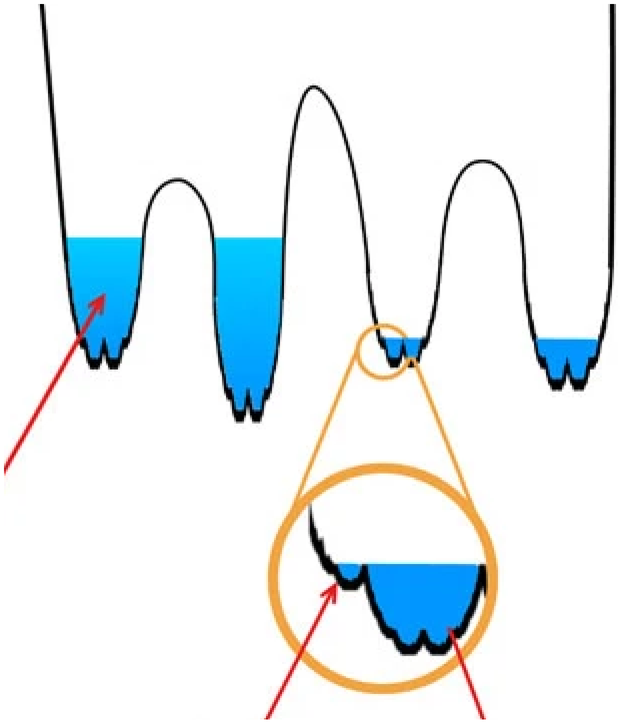


In order to be expressive enough, the RBM must describe all possible levels of similarity

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

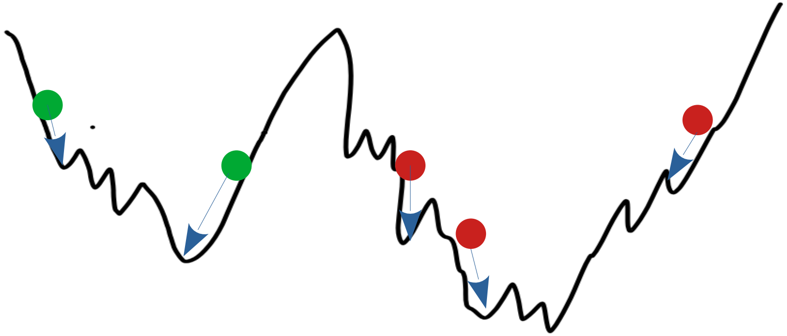


Data has a hierarchical organization



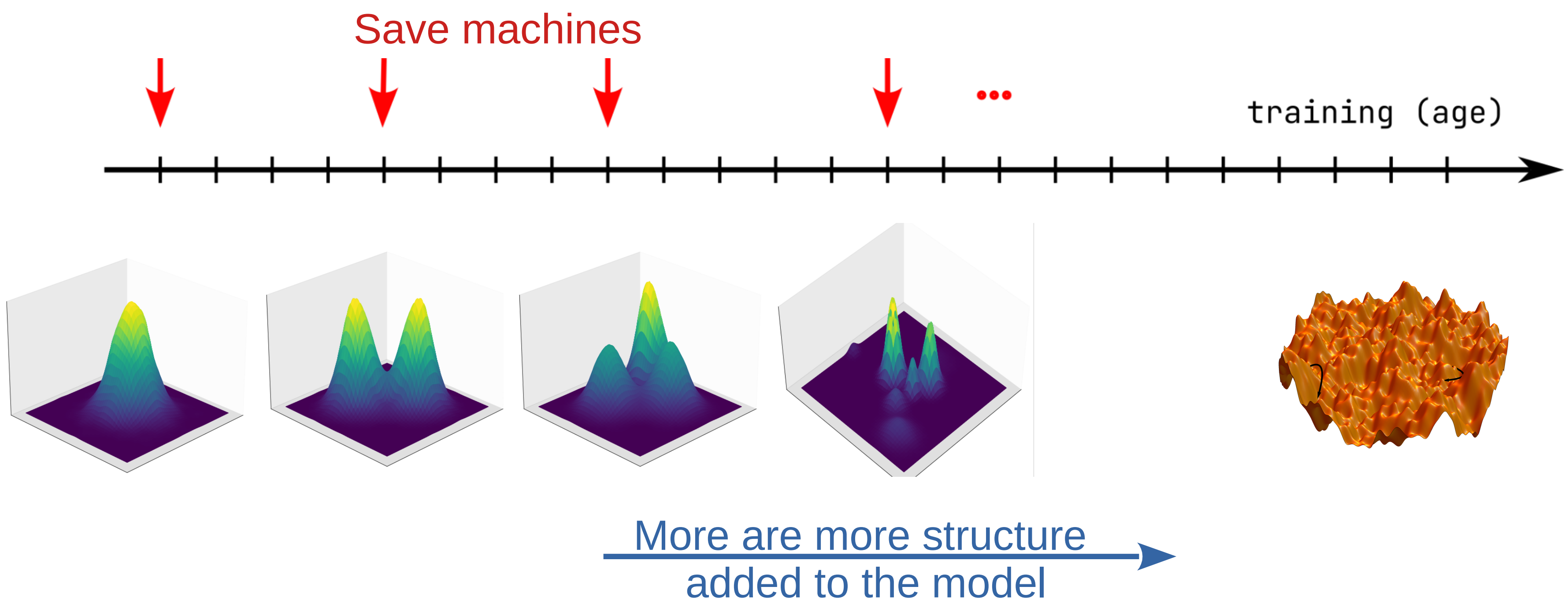
In order to be expressive enough, the RBM must describe all possible levels of similarity

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

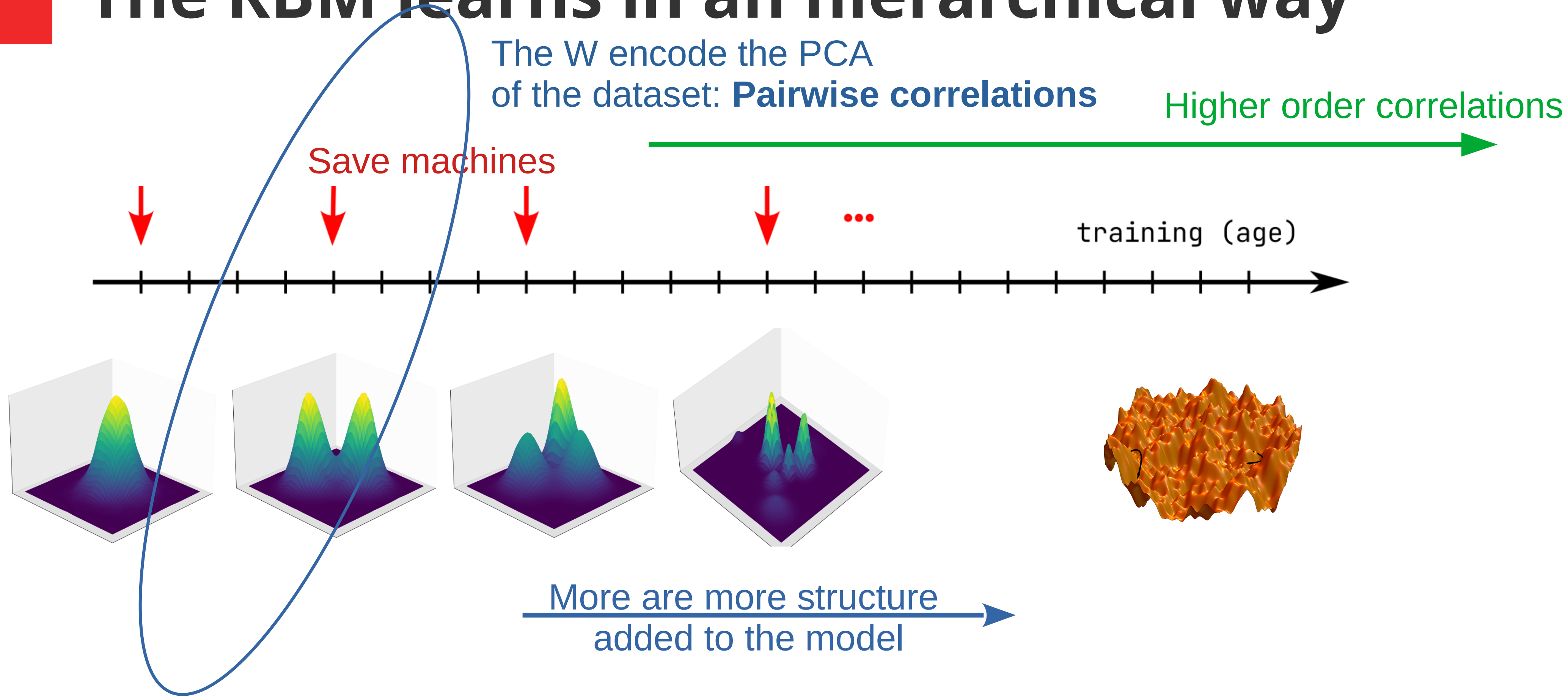


How do we detect larger basins?

The RBM learns in an hierarchical way



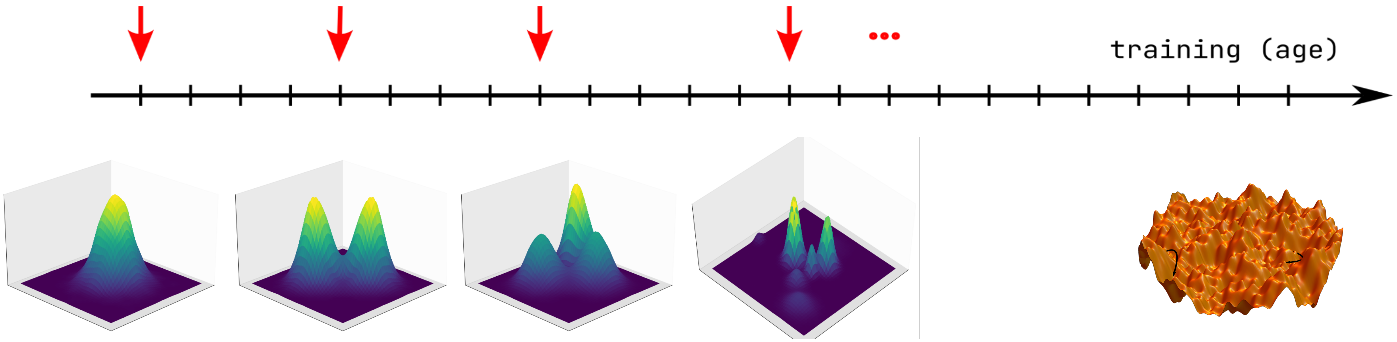
The RBM learns in an hierarchical way

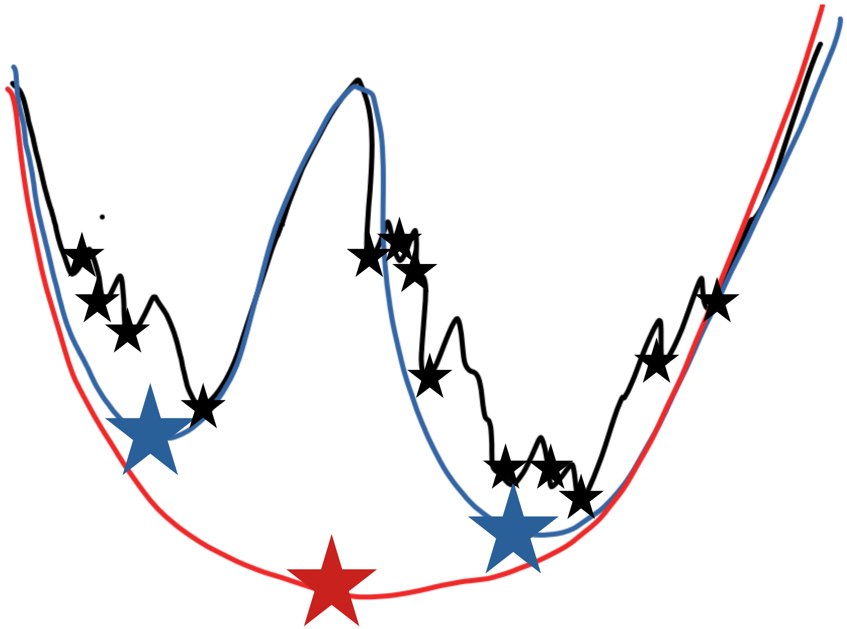


* Decelle, Fissore and Furtlehner, *Spectral dynamics of learning in restricted boltzmann machines* (2017)

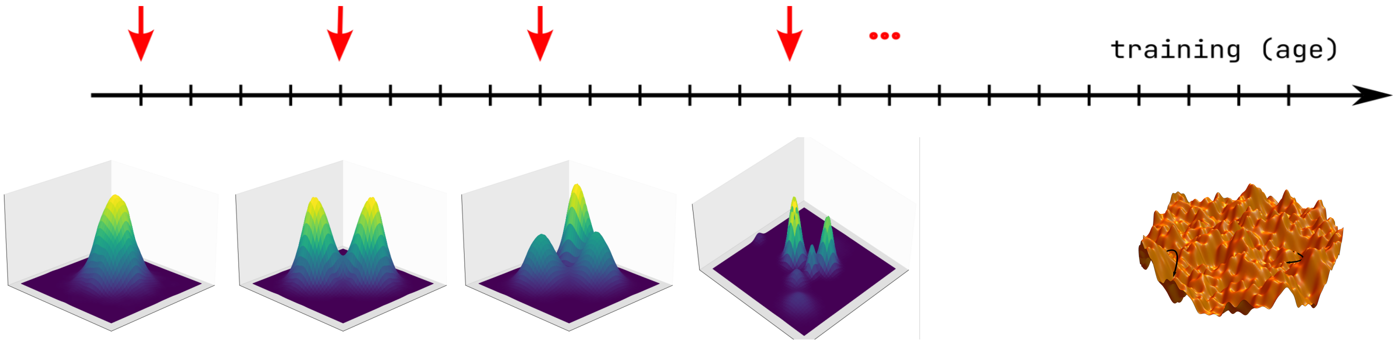
* Decelle, & Furtlehner, *Restricted Boltzmann machine: Recent advances and mean-field theory* (2021)

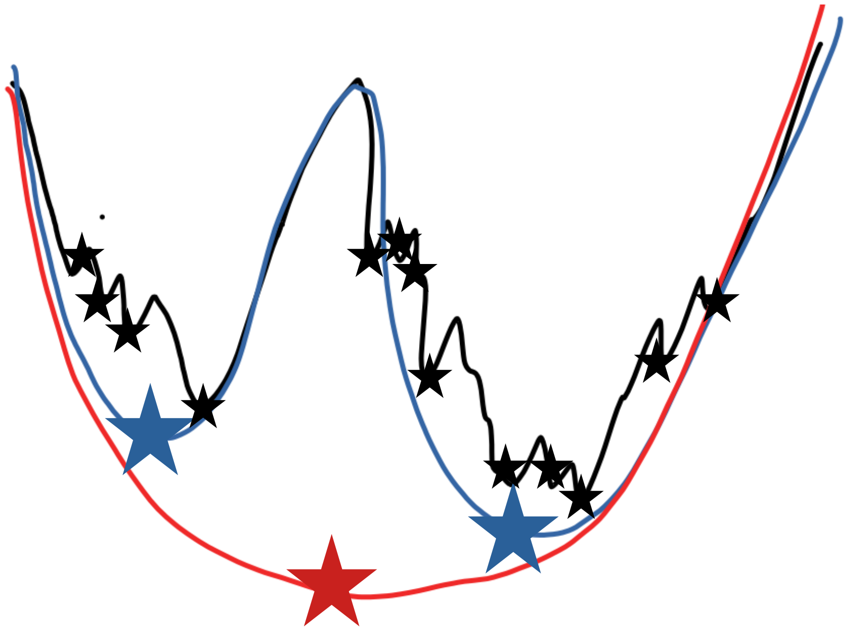
Hierarchical Clustering

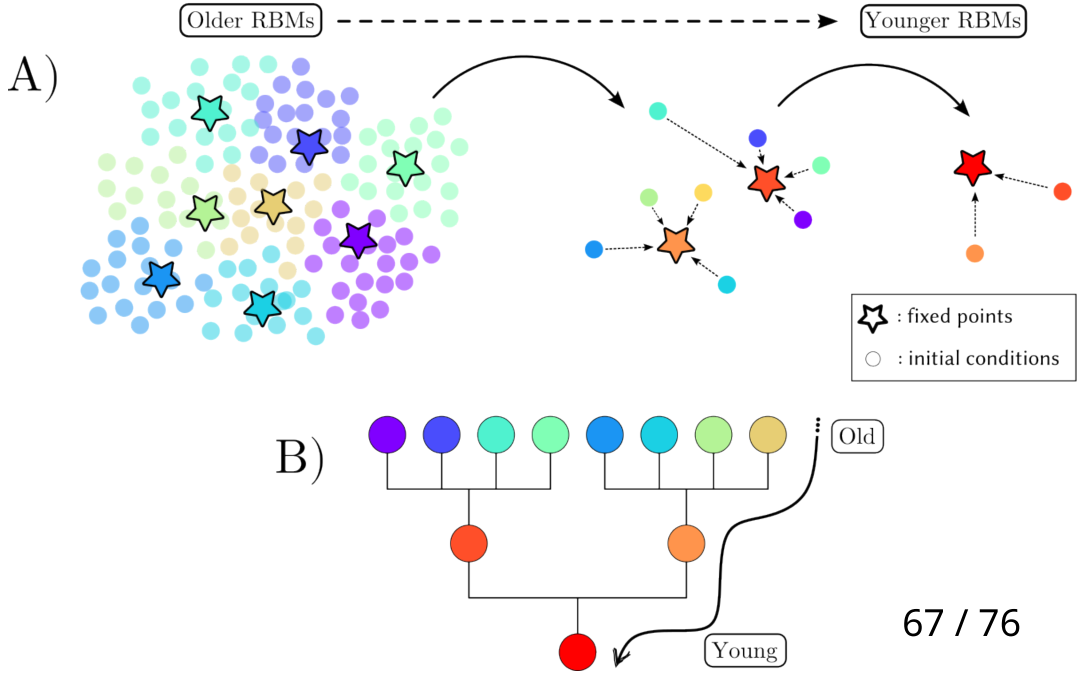




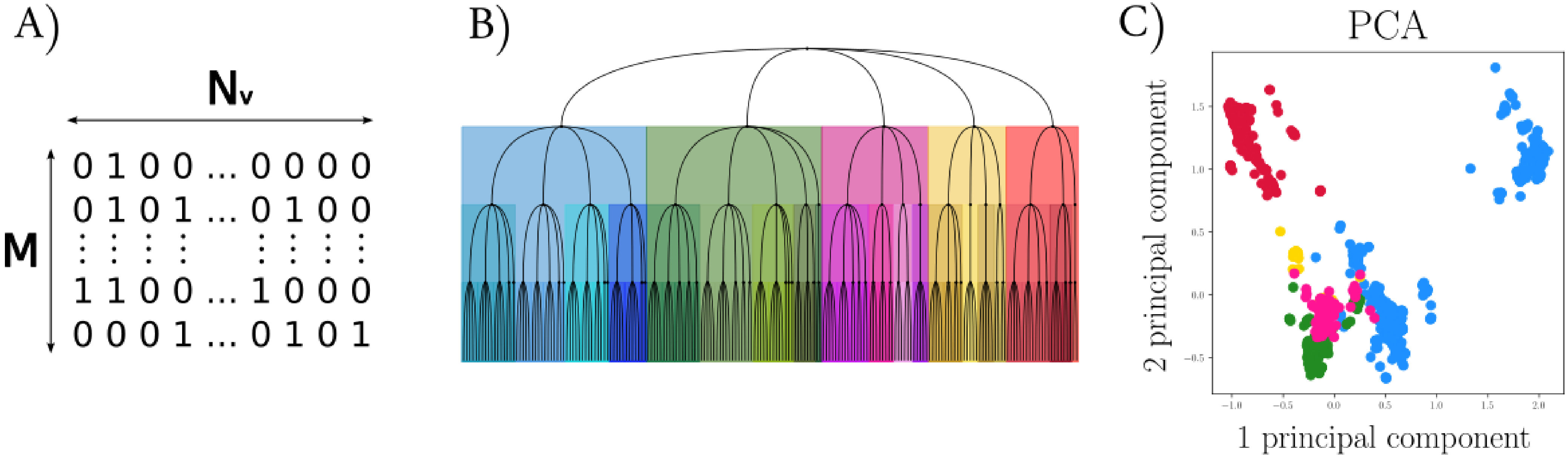
Hierarchical Clustering



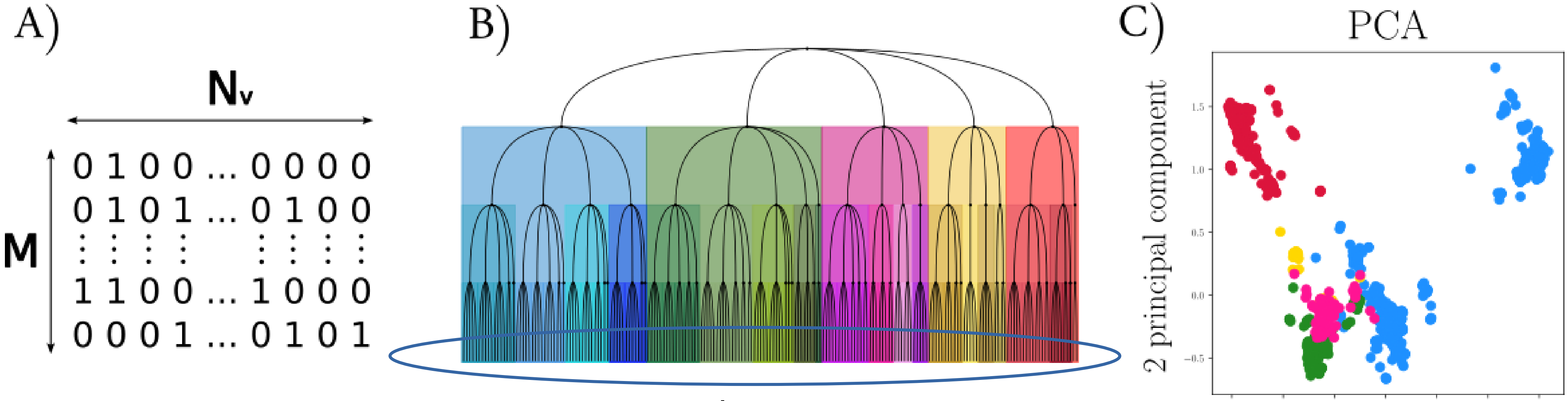




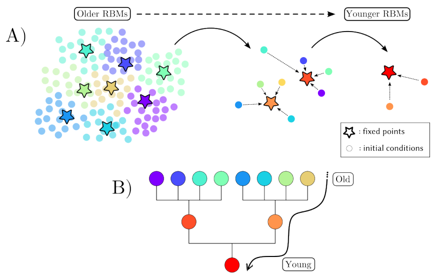
Example: synthetic evolutionary data



Example: synthetic evolutionary data

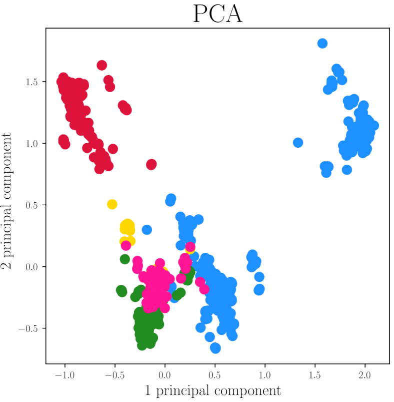


Train a RBM

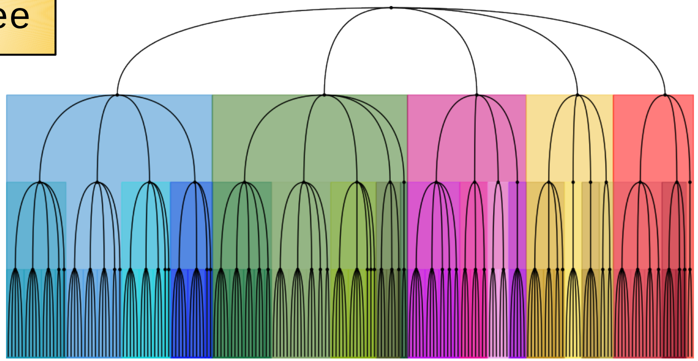


Build a tree
Using machines saved during the training

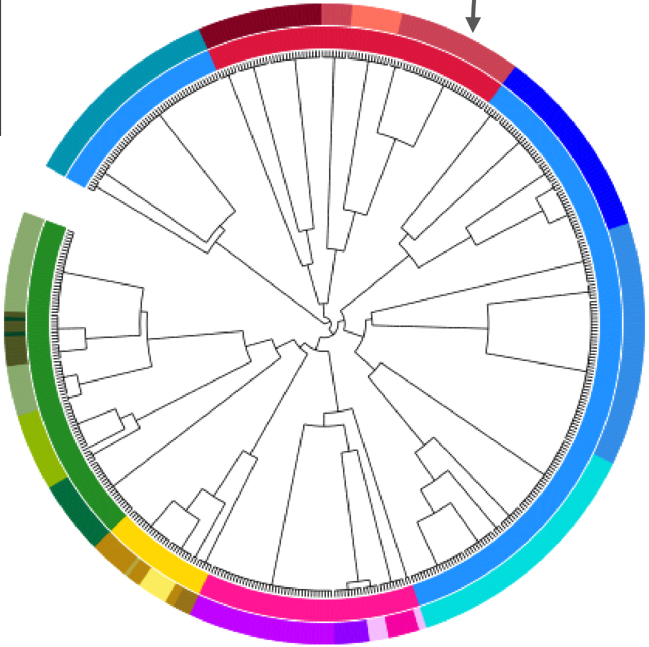
Synthetic data



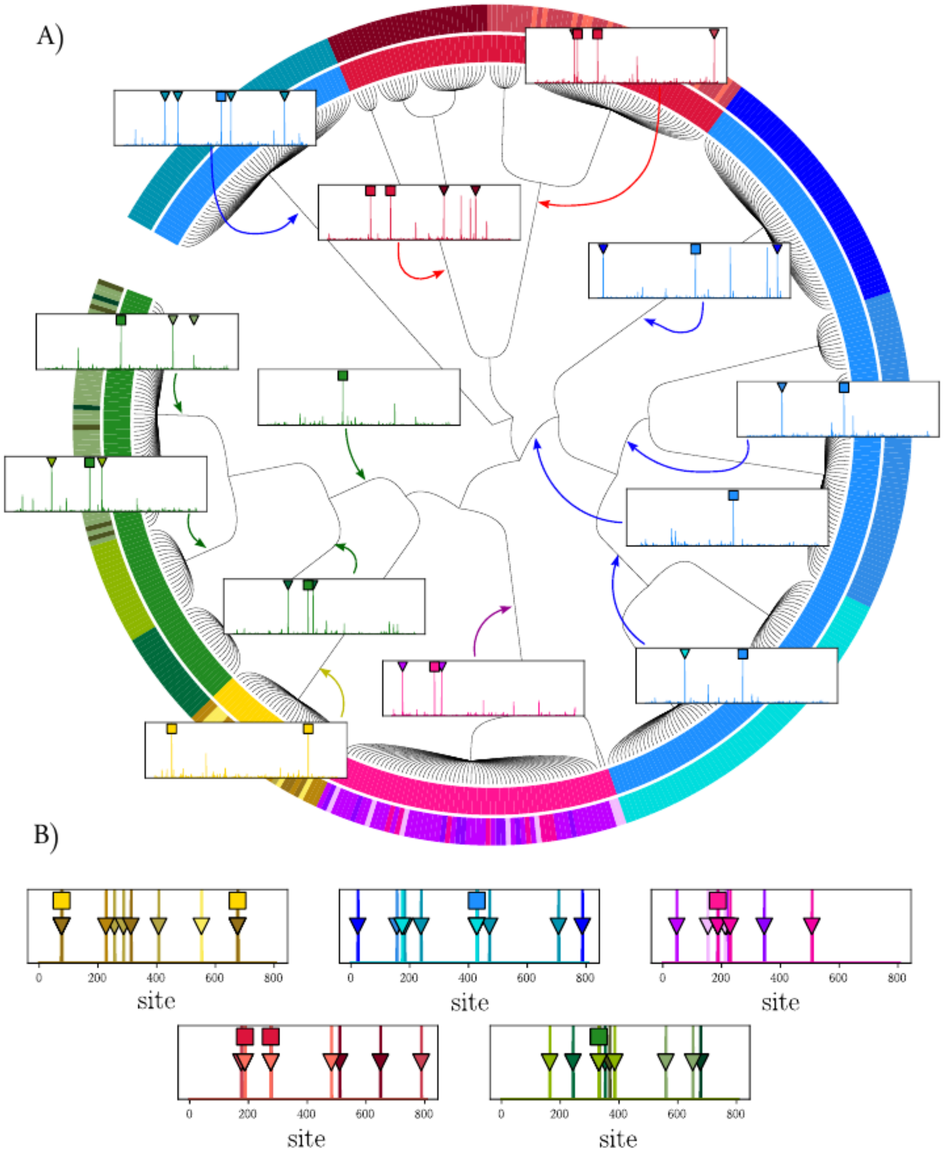
Real tree



Reconstruction

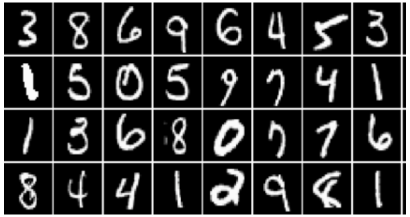


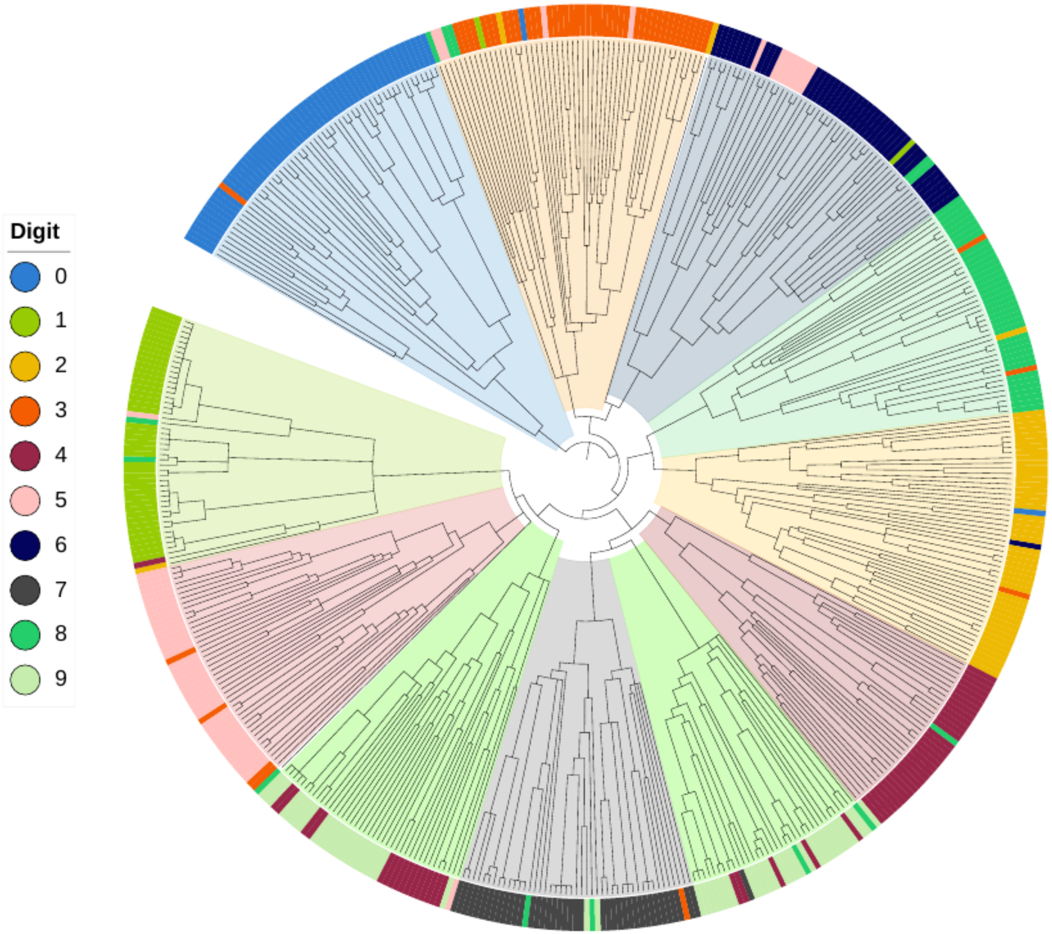
Synthetic data



Hierarchical Clustering

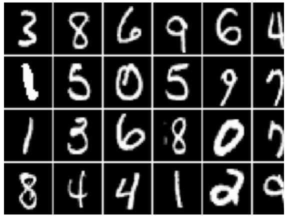
MNIST data

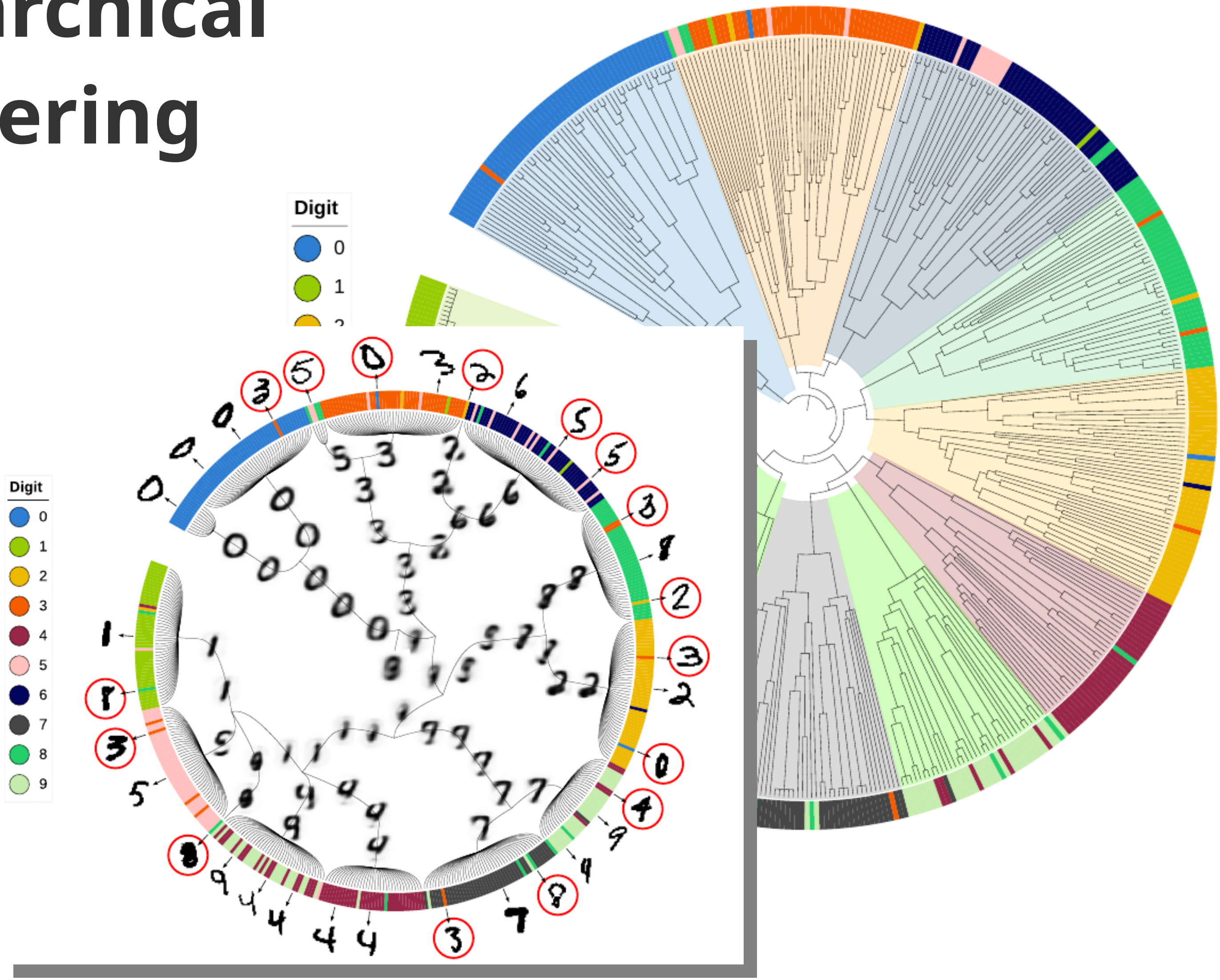




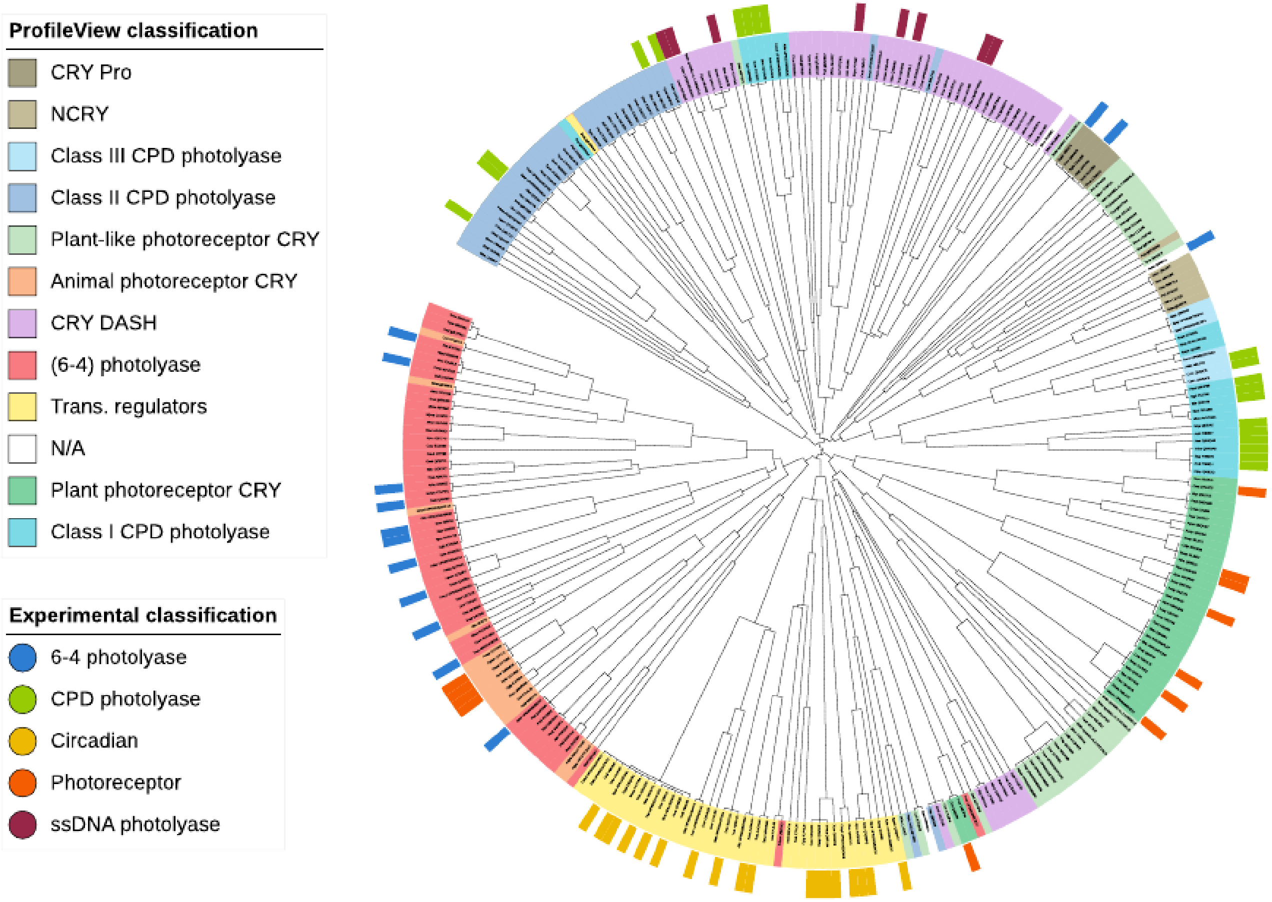
Hierarchical Clustering

MNIST data



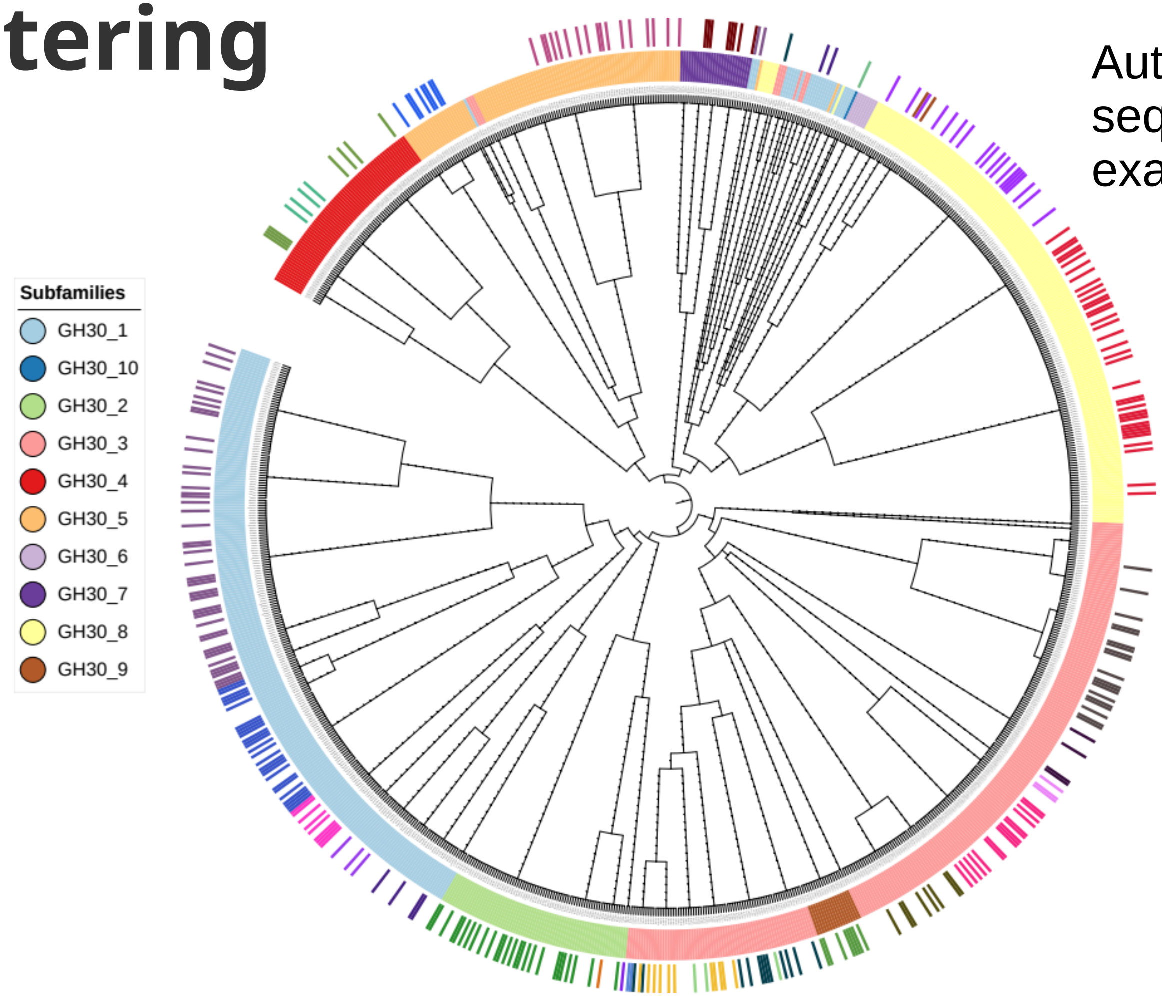


Protein function classification



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