Learning protein constitutive motifs from sequence data with restricted Boltzmann machines

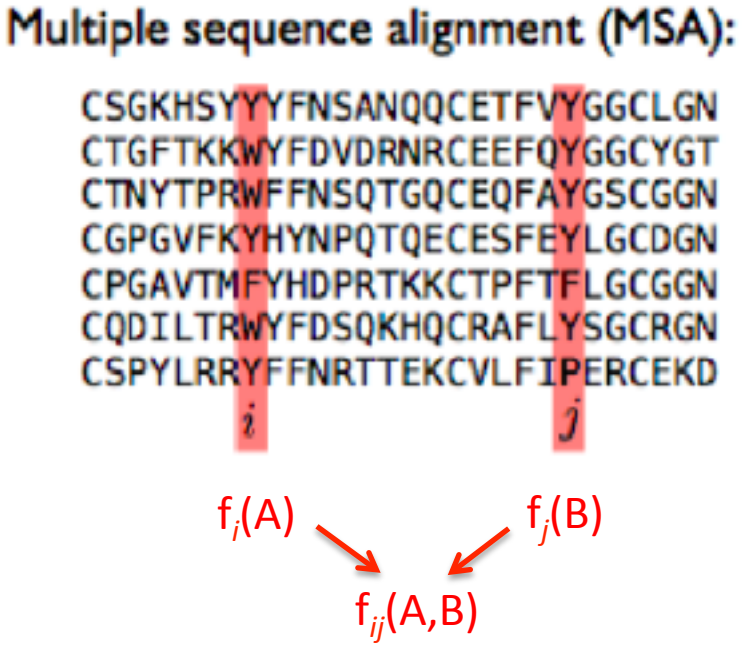
R. Monasson, Theoretical Physics Lab., Ecole Normale Supérieure & CNRS, Paris

In collaboration with:

S. Cocco, Statistical Physics Lab., ENS & CNRS
J. Tubiana, Theoretical Physics Lab., ENS

See: manuscript arxiv:1803.08718

Covariation in sequence alignments

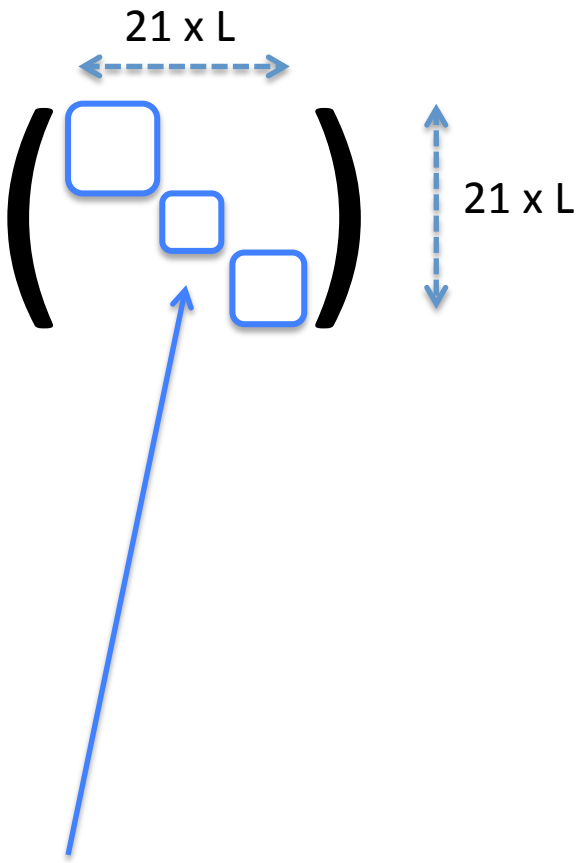


Pairwise Correlation Matrix: $C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)\, f_j(B)$

NB: proteins: tens to hundreds amino acids ➔ nb of entries: millions to hundreds of millions
alignments: thousands to tens of thousands of sequences

What to do with the pairwise correlation matrix?

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)\, f_j(B) =$$



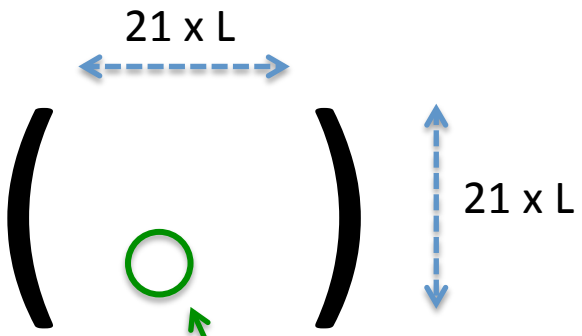
What can we do with this matrix?

1. Look for collective modes:
   - Extract dominant directions (Principal Component Analysis)
   - Clustering of sites in low-dimensional space reveals groups of functionally co-evolving residues

[Russ et al, Nature 2005]
[Halabi, Rivoire, Leibler, Ranganathan, Cell 2009]
[De Juan, Pazos, Valencia, Nat Rev Gen 2013]

What to do with the pairwise correlation matrix?

$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)\, f_j(B) =$

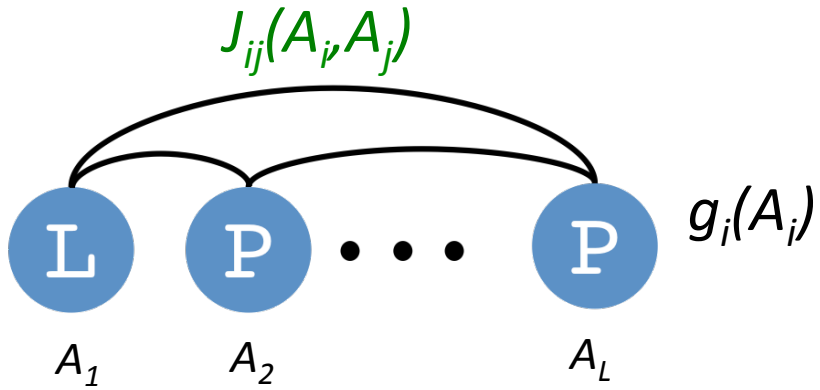


What can we do with this matrix?

2. Look for interactions reproducing correlations:

Correlations are mediated by paths of direct interactions

[Lapedes et al, unpublished 2001; Weigt et al, PNAS 2009]



Probabilistic score of sequence:

$$\log P(A) = \sum_i g_i(A_i) + \sum_{i<j} J_{ij}(A_i, A_j)$$

indicative of structure

Hereafter, a unifying method to learn motifs controlling structural, functional, evolutionary properties of proteins from sequence data

- How it works: Build **representations** of sequences
  (idea coming from unsupervised learning)

- What it gives: Application to protein domains

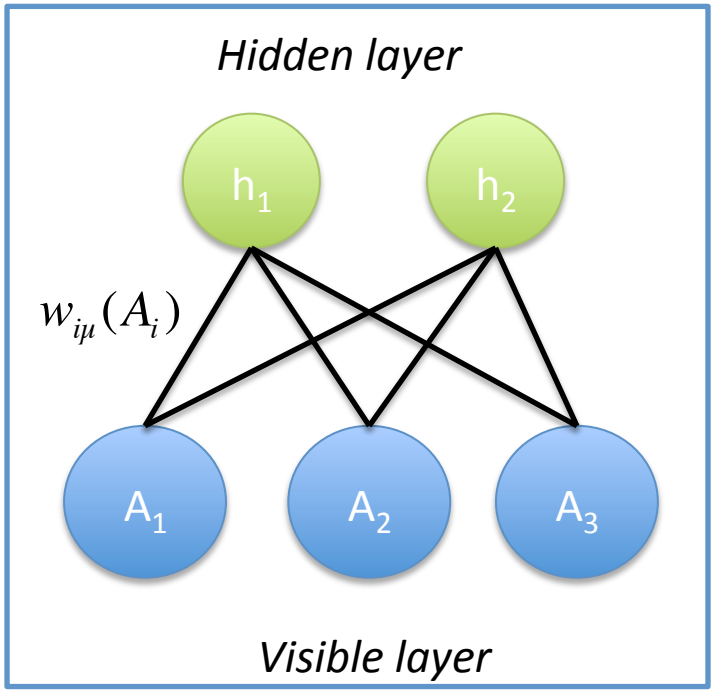
- Why it works: Control of the operation point of the machine-learning method

Restricted Boltzmann Machines

- **Graphical model** constituted by two sets of random variables that are coupled together.

$$\log P(A,h) = \sum_i g_i(A_i) + \sum_{i,\mu} w_{i\mu}(A_i)\, h_\mu - \sum_\mu U_\mu(h_\mu)$$

gives *P(A)* after integration over h's…



Ackley, Hinton, Sejnowsky 1985
Smolensky 1986

Restricted Boltzmann Machines

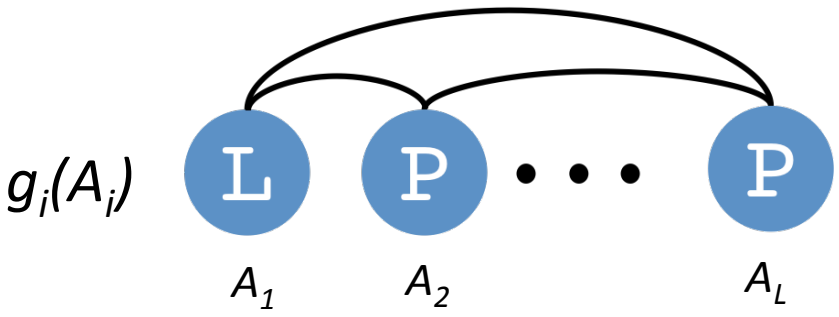
- **Graphical model** constituted by two sets of random variables that are coupled together.

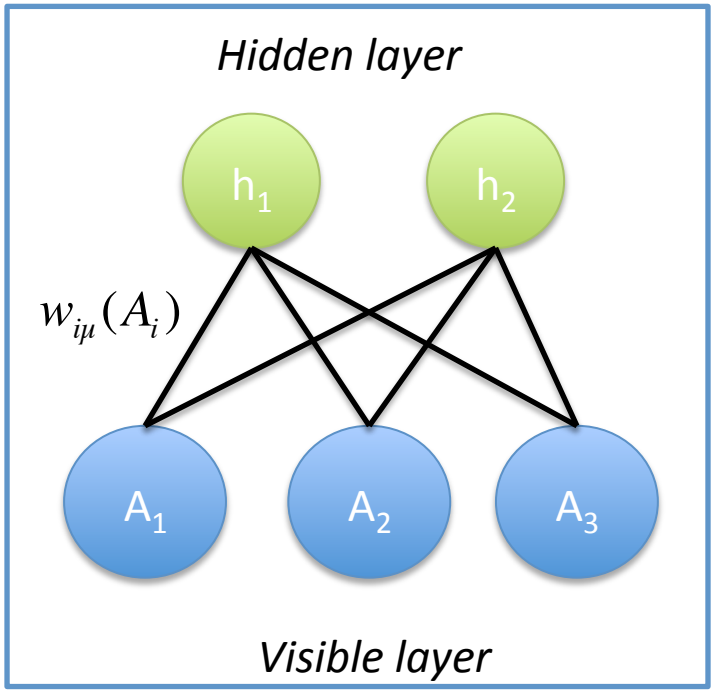
$$\log P(A,h) = \sum_i g_i(A_i) + \sum_{i,\mu} w_{i\mu}(A_i)\ h_\mu - \sum_\mu U_\mu(h_\mu)$$

gives *P(A)* after integration over h's…

Quadratic U(h): identical to Potts model!!

$$J_{ij}(A_i, A_j) = \sum_\mu w_{i\mu}(A_i)\ w_{j\mu}(A_j)$$



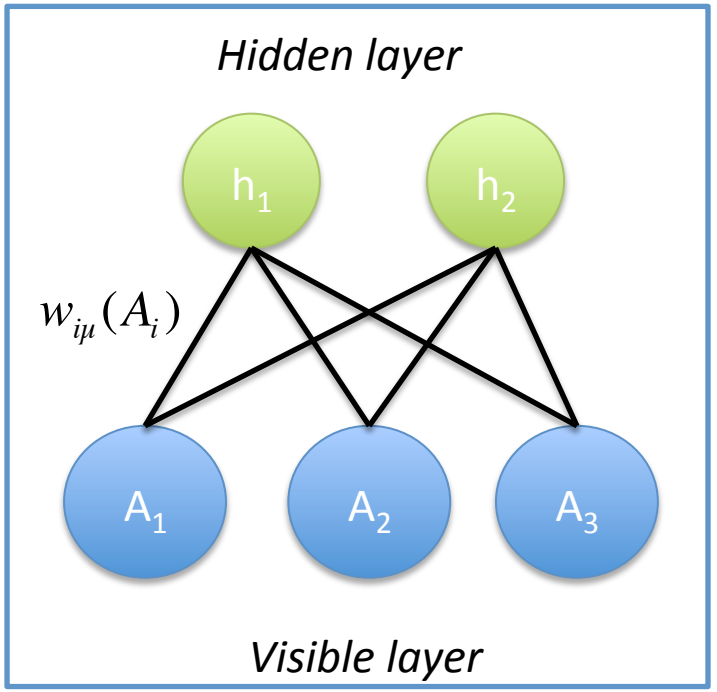


Ackley, Hinton, Sejnowsky 1985
Smolensky 1986

Restricted Boltzmann Machines

- **Graphical model** constituted by two sets of random variables that are coupled together.

$$\log P(A,h) = \sum_i g_i(A_i) + \sum_{i,\mu} w_{i\mu}(A_i)\, h_\mu - \sum_\mu U_\mu(h_\mu)$$



- Joint distribution of *A,h* define

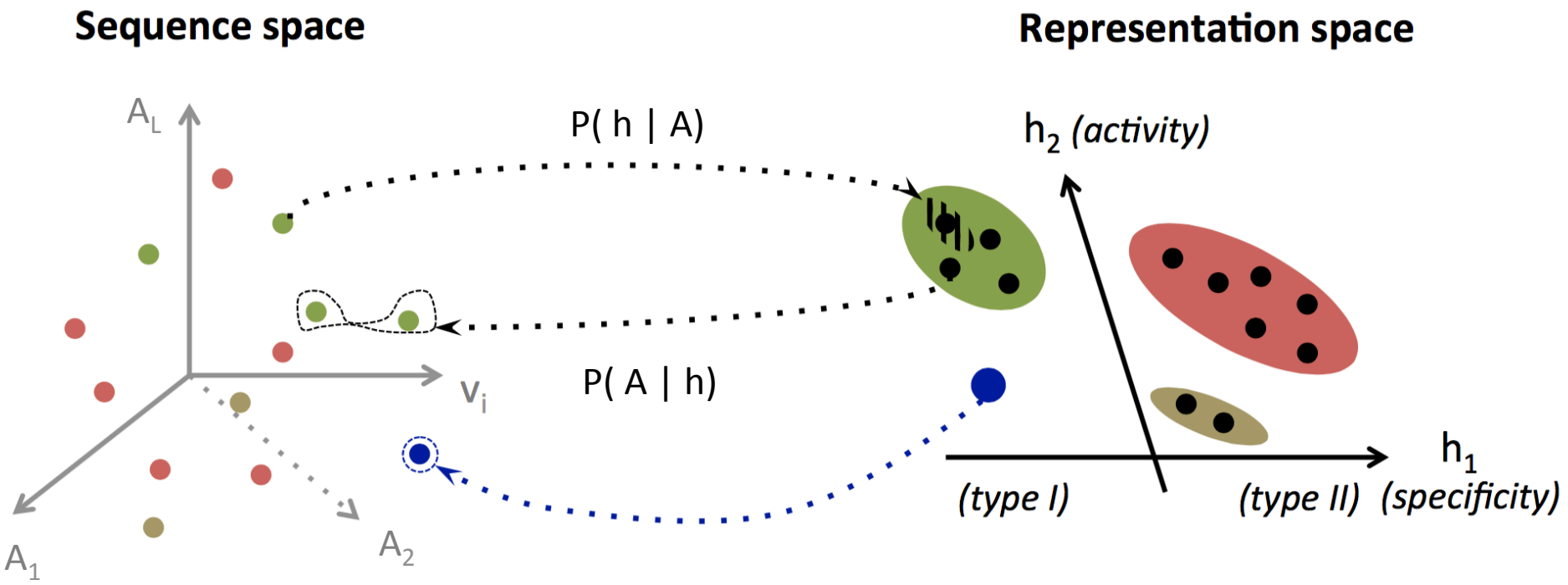
$$\log P(A), \quad \log P(h|A), \quad \log P(A|h)$$

maximized over sequence data set

Extract latent factors (representation) from sequence

Design sequence from representation

High-dimensional representations of protein sequences

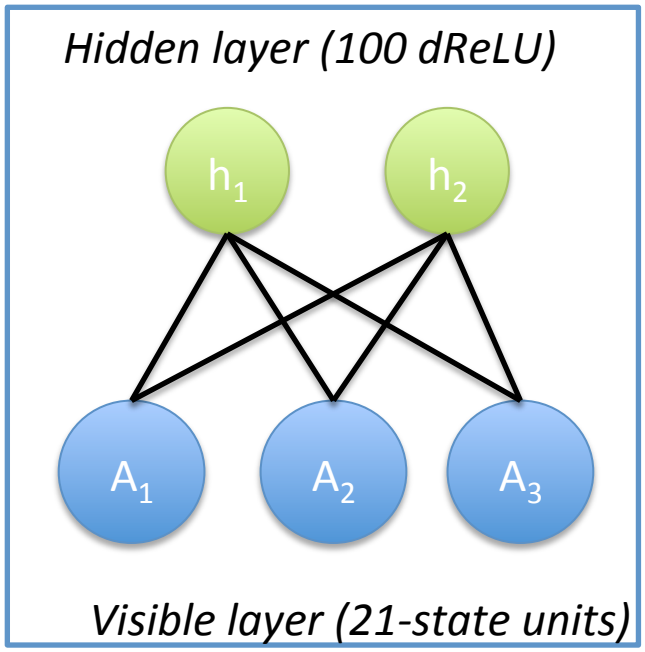


- RBM extract high-D representations of (common inputs to) sequences
- Representations are useful (to design « good » sequences) …
- … and, hopefully, biologically meaningful (structure, function, history)

➔ Practical implementation of genotype-to-phenotype relation

Applications of RBM to protein sequence data

WW domain
(PFAM PF00397)

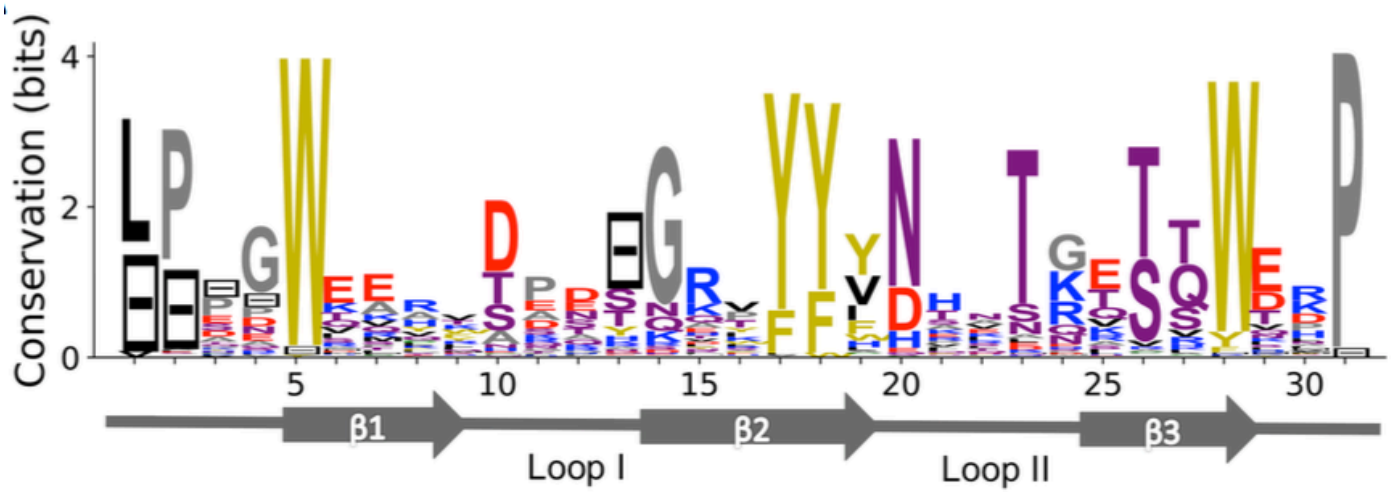


Small domain with ~30 a.a.

Can be done for much longer proteins, e.g.
Trypsin (protease) with ~220 a.a.
HSP70 (chaperone) with ~600 a.a.

WW domain

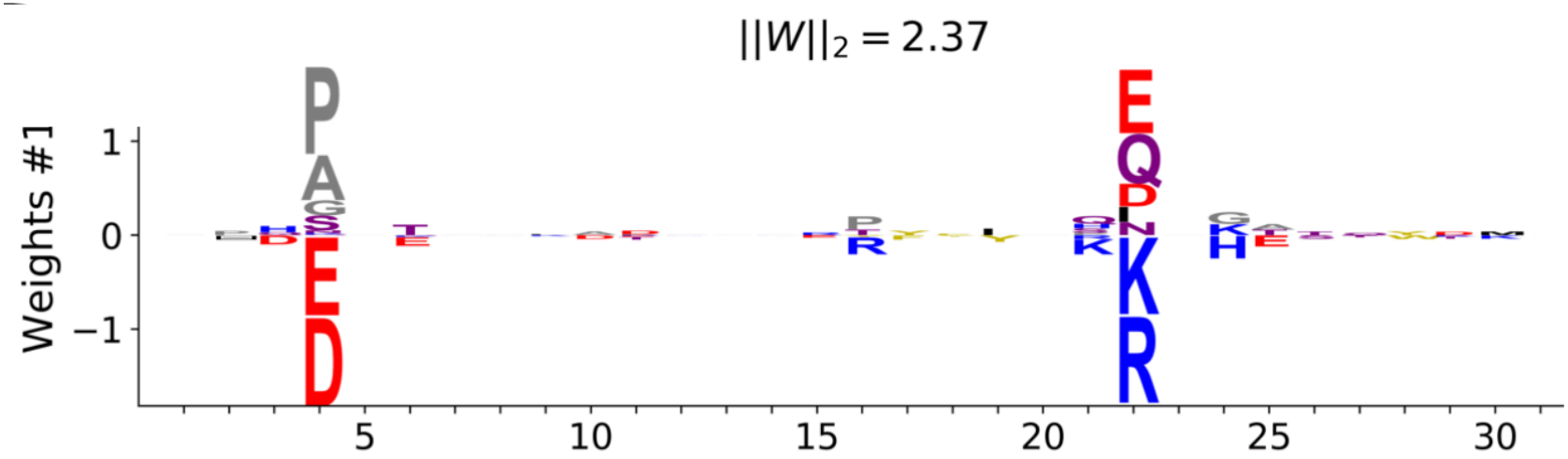
= short binding domain involved in eukaryotic signalling proteins
folds into 3-stranded antiparallel beta sheet



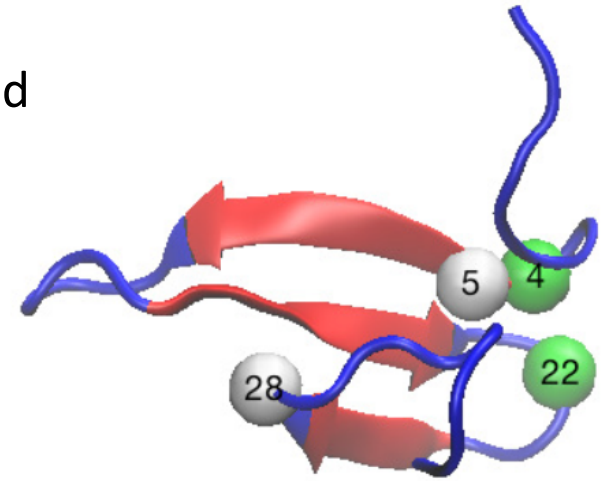
Binds to four different types of Proline (P) rich ligands:

- Type I: PPXY, Y = Tyrosine (aromatic) , X = any residue
- Type II: PPLP, L = Leucine
- Type III: PR rich peptide, R = arginine
- Type IV: p(S/T)P, phosphorylated serine/threonine

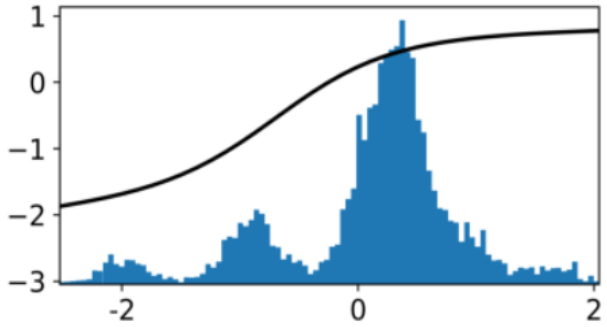
Weights may correspond to contacts



3D fold

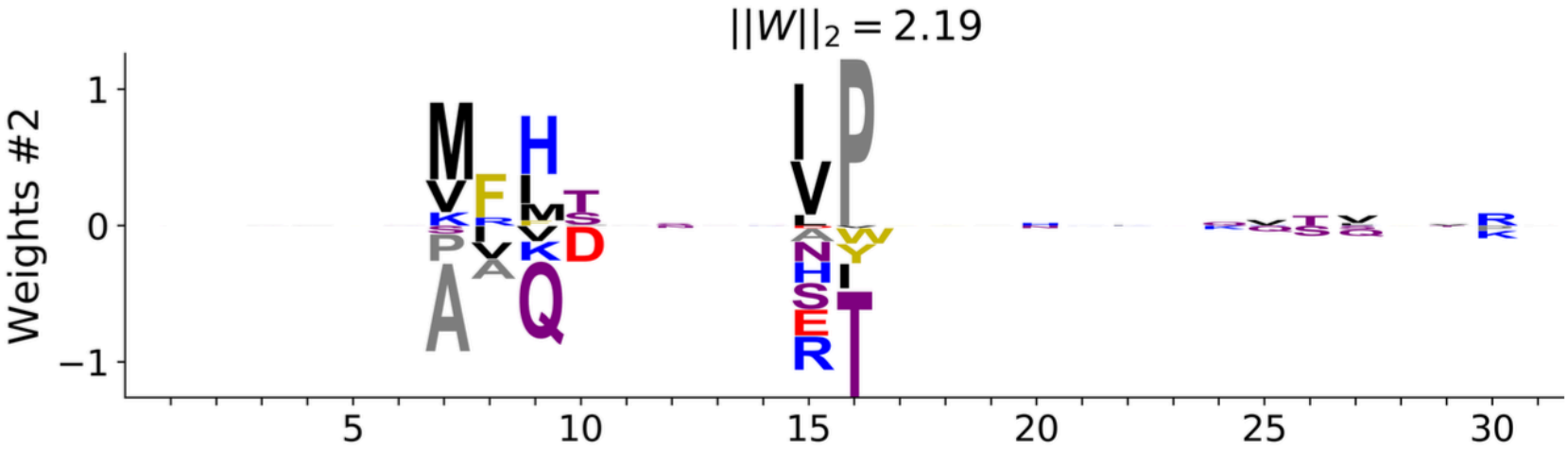


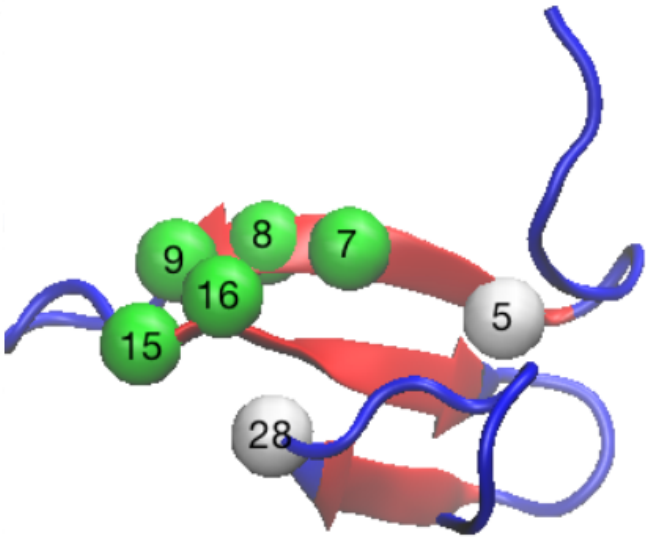
Distribution across MSA

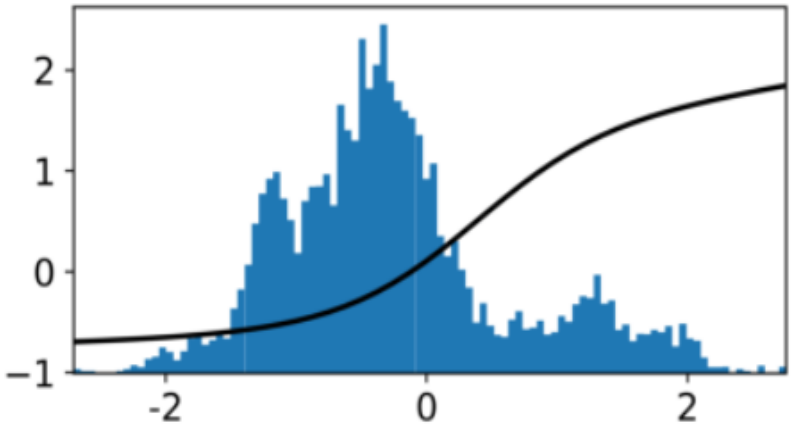


$$I_\mu(A) = \sum_i w_{i\mu}(A_i)$$

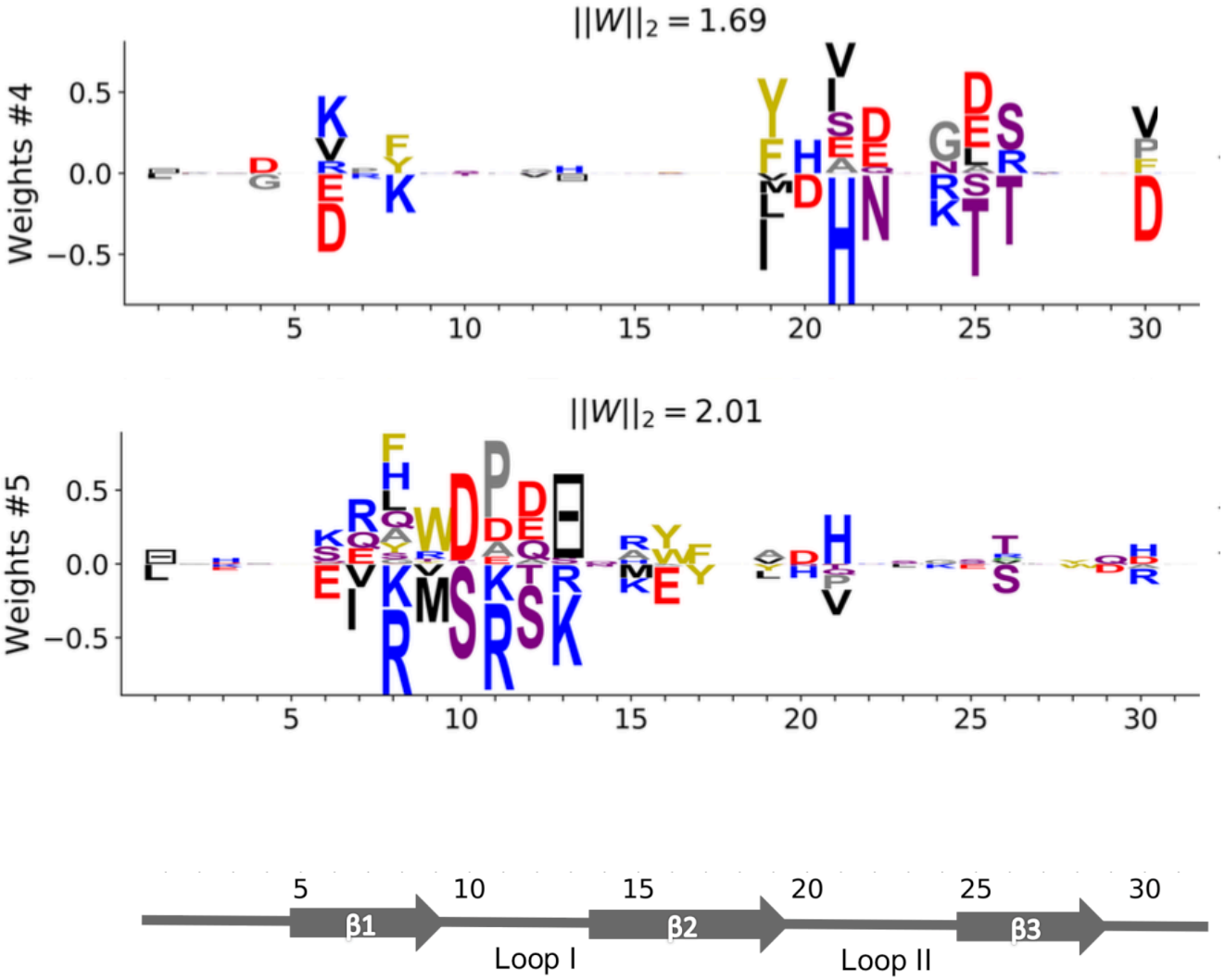
Weights may correspond to structural modes

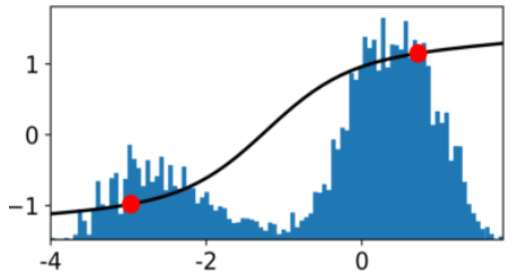


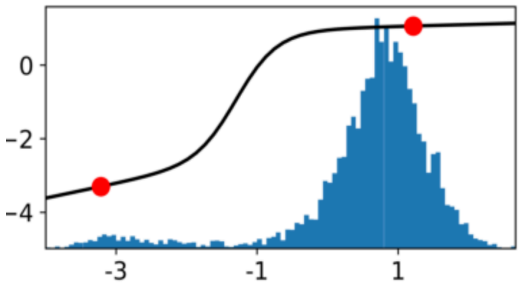




Weights may describe functional specificity

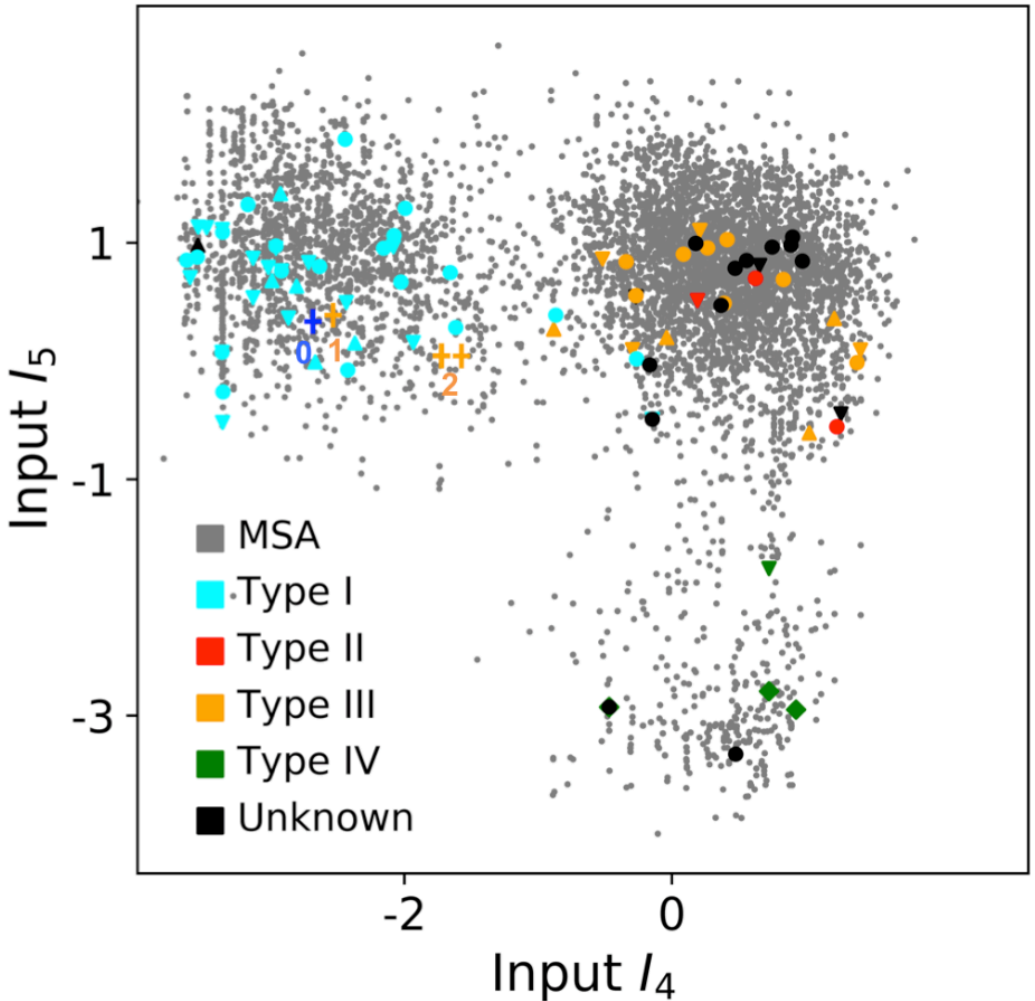






Weights may describe functional specificity

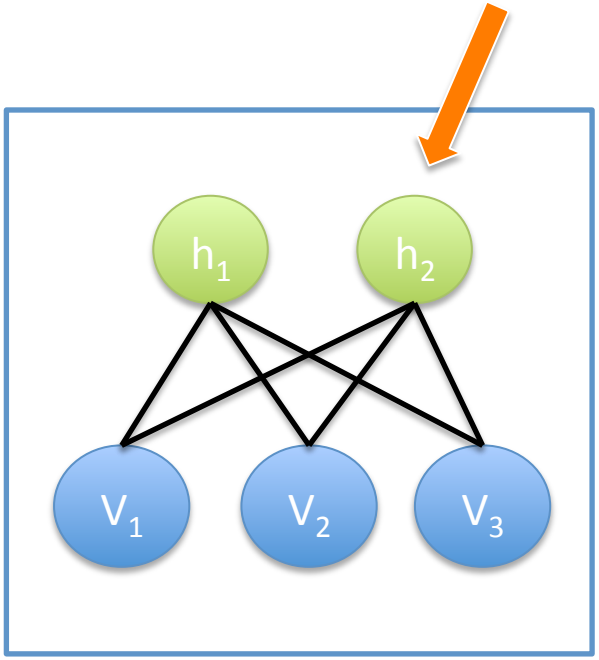
- Type I: PPXY, Y = Tyrosine (aromatic) , X = any residue
- Type II: PPLP, L = Leucine
- Type III: PR rich peptide, R = arginine
- Type IV: p(S/T)P, phosphorylated serine/threonine

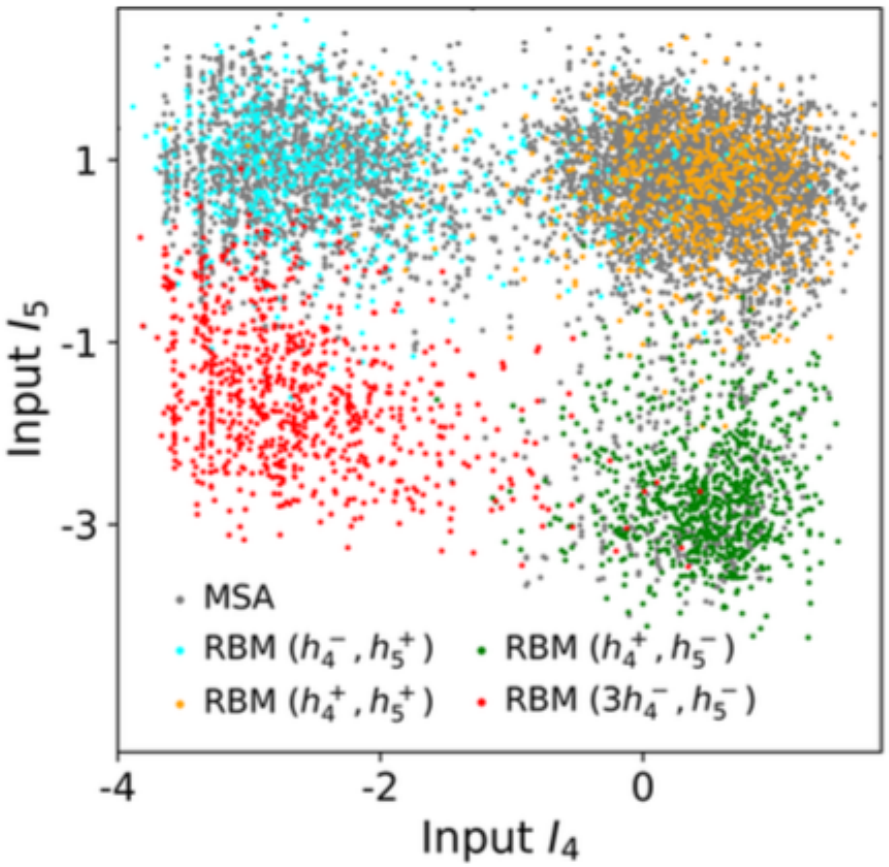


Ingham et al. Molecular and Cell Biology 2005
Jager et al, PNAS 2006

Conditional design of sequences

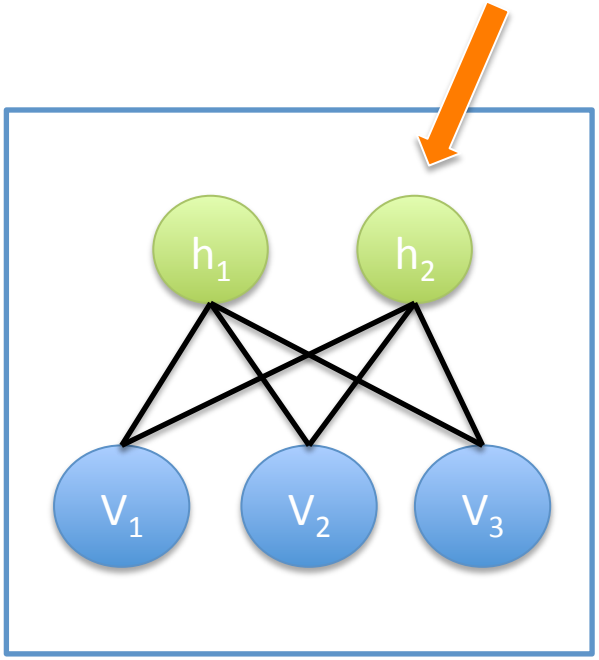
Once we understand what hidden units code for, we may bias sampling …

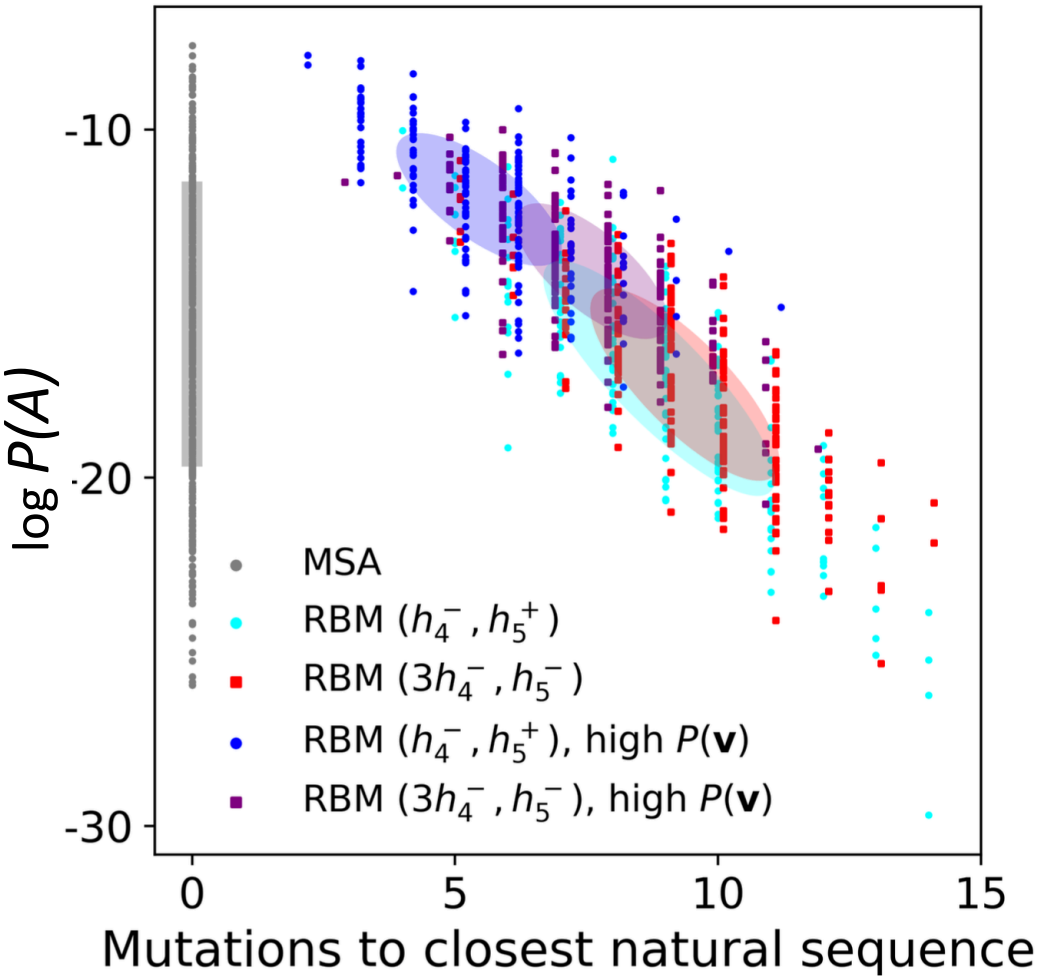




RBM are able to generate sequences in a restricted portion of the sequence space, even unexplored by « natural » sequences !

Conditional design of sequences





Generated sequences have high probabilities, and are far away from natural sequences !

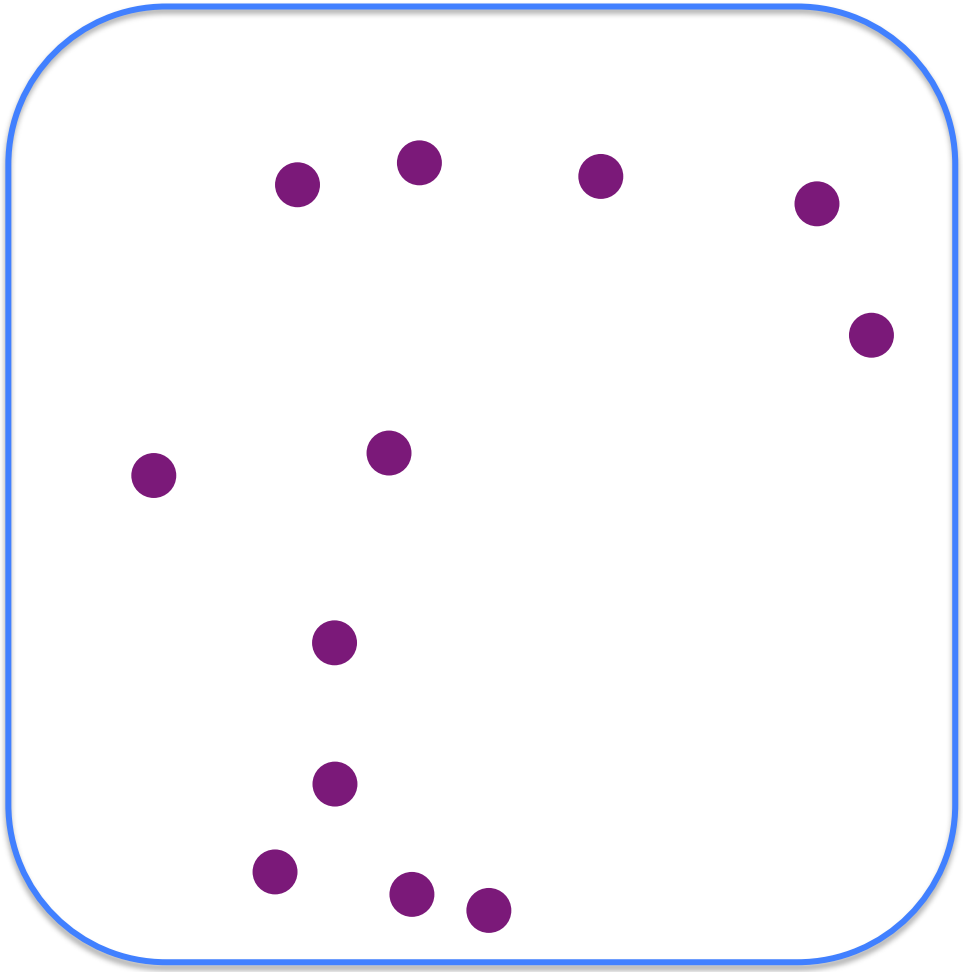
Approach benchmarked on synthetic protein models

Ongoing experiments …

Why are RBM extracting useful/interpretable representations?

The problem:

Find probability distribution from very few samples

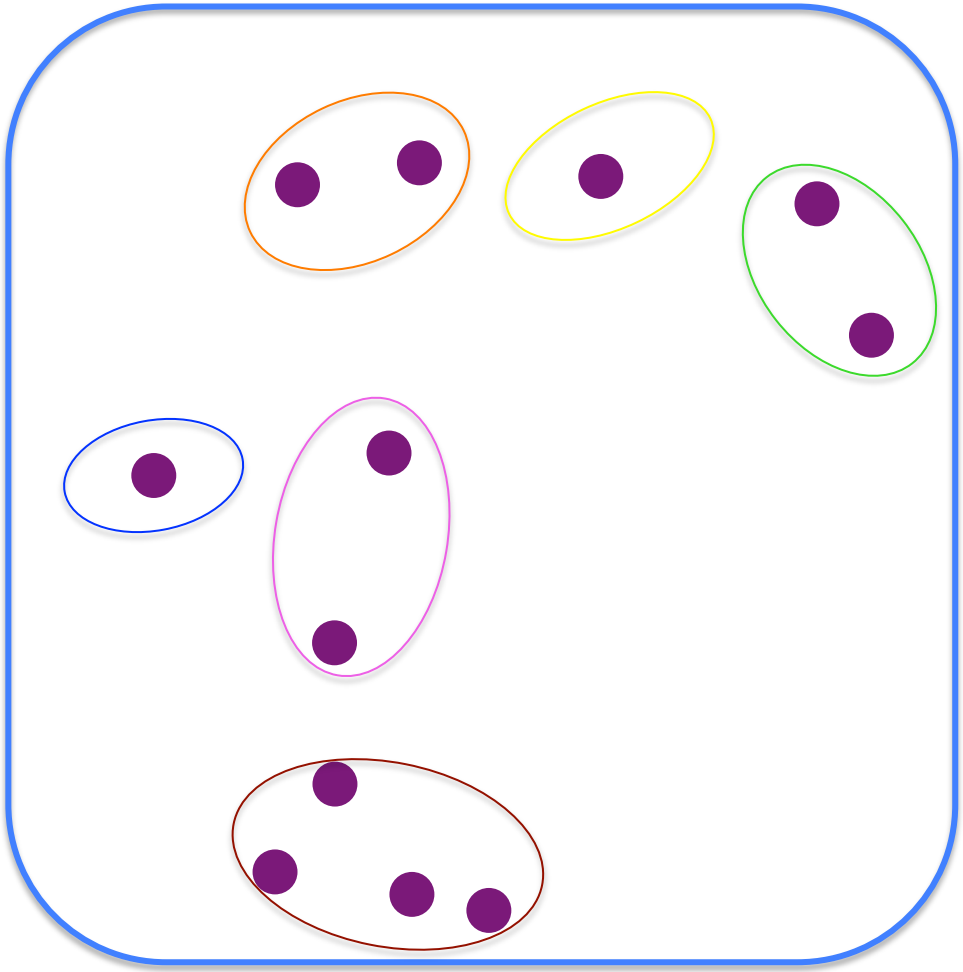


Sequence space

Why are RBM extracting useful/interpretable representations?

Mixture of local models :

Each hidden unit sees and codes for a patch in sequence space



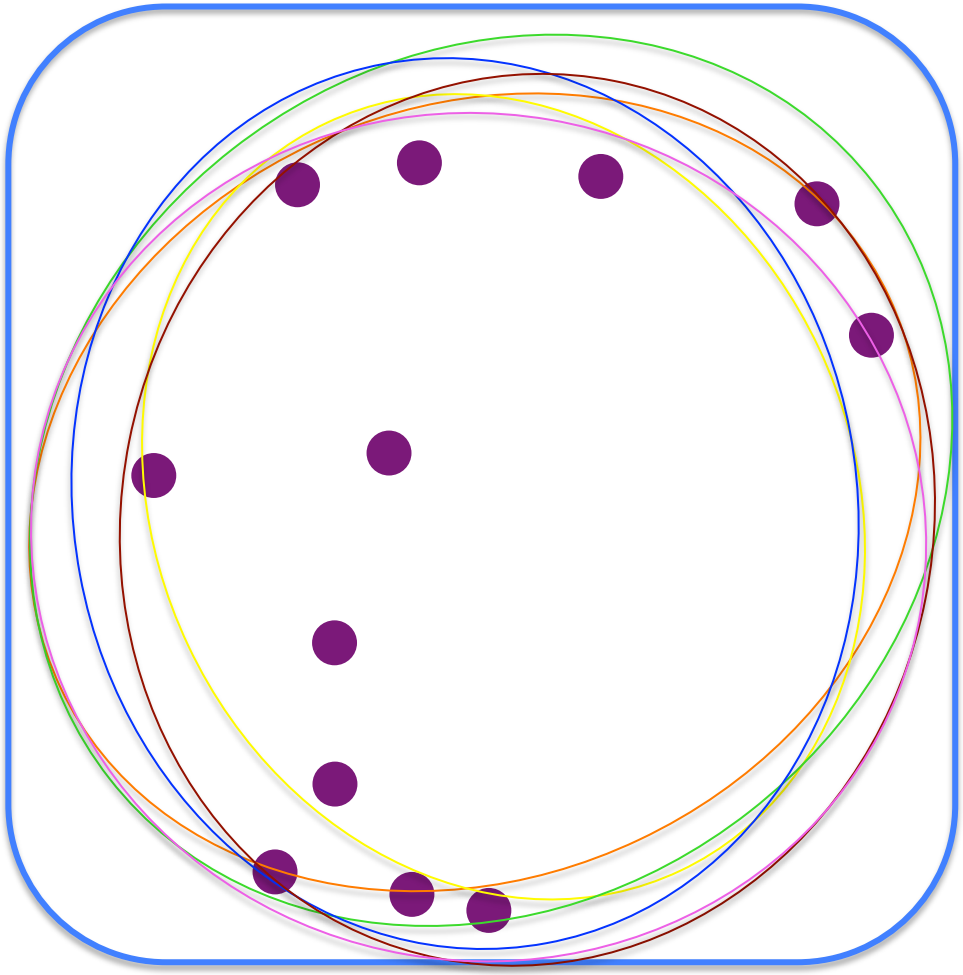
Sequence space

Why are RBM extracting useful/interpretable representations?

Entangled model:

All or almost all hidden
units active at any position in sequence space

Non interpretable representations …

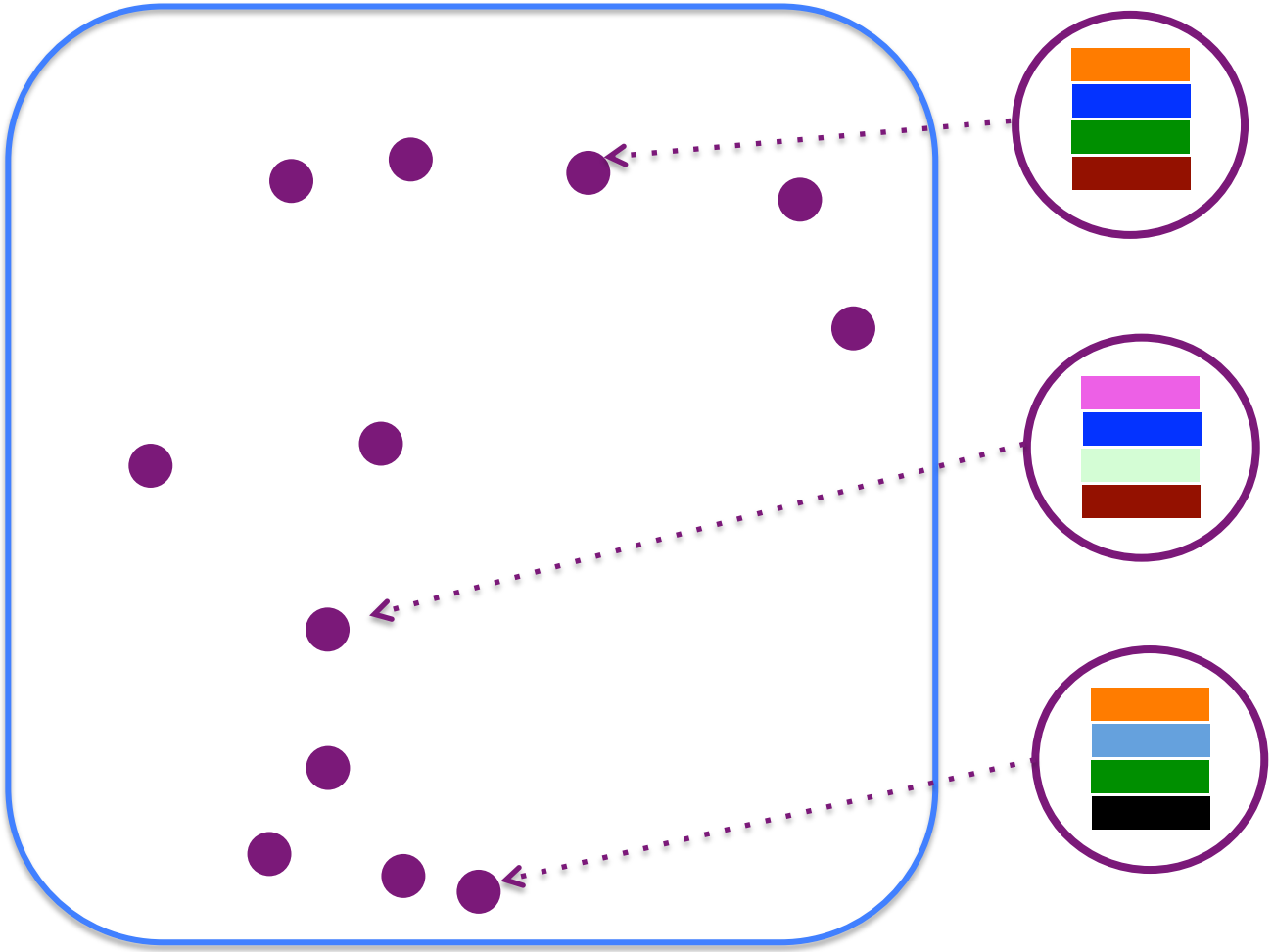


Sequence space

Why are RBM extracting useful/interpretable representations?

Decomposition into constitutive features:

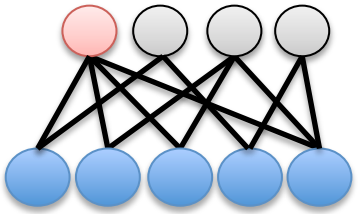
Each hidden unit codes for an invariant feature; sequences are obtained by combinatorial composition of features



Sequence space

The three representational regimes of RBM

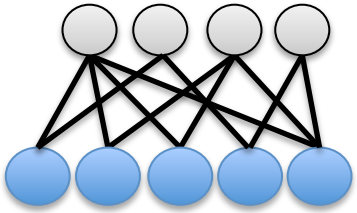
"Mixture of local models" regime



- One hidden unit very active
- Corresponding weights define local prototype
- Unable to extract invariances

More hidden units per visible unit

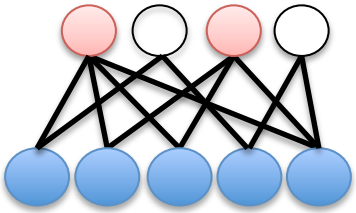
"Globally Distributed" regime



- All hidden units active
- Visible configurations are complex mixtures
- Not Interpretable

Increasing sparsity

Non quadratic hidden-unit potentials

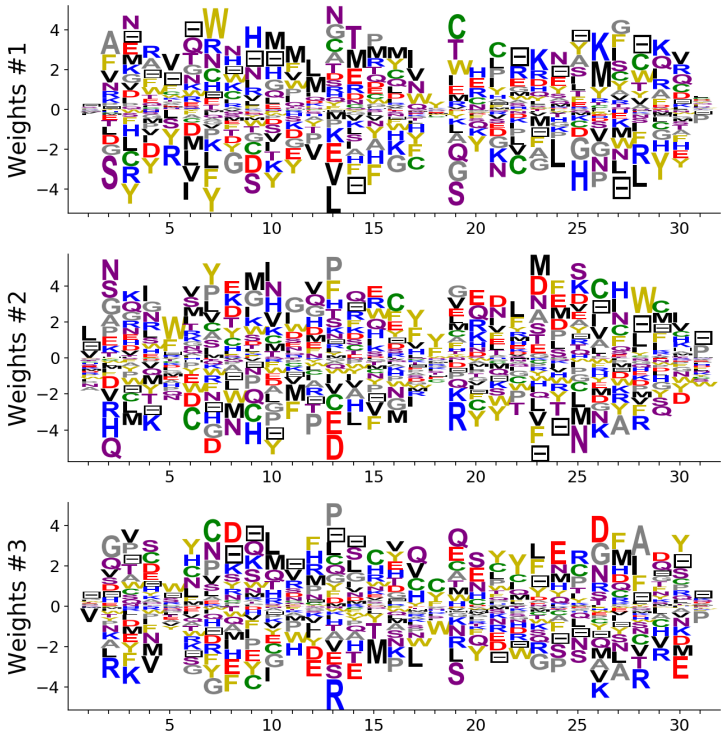


"Compositional" regime

- Multiple hidden units very active
- Corresponding weights define features composing visible configurations
- Possibly interpretable

Tubiana, Monasson, PRL 2017

Driving RBM to the compositional phase

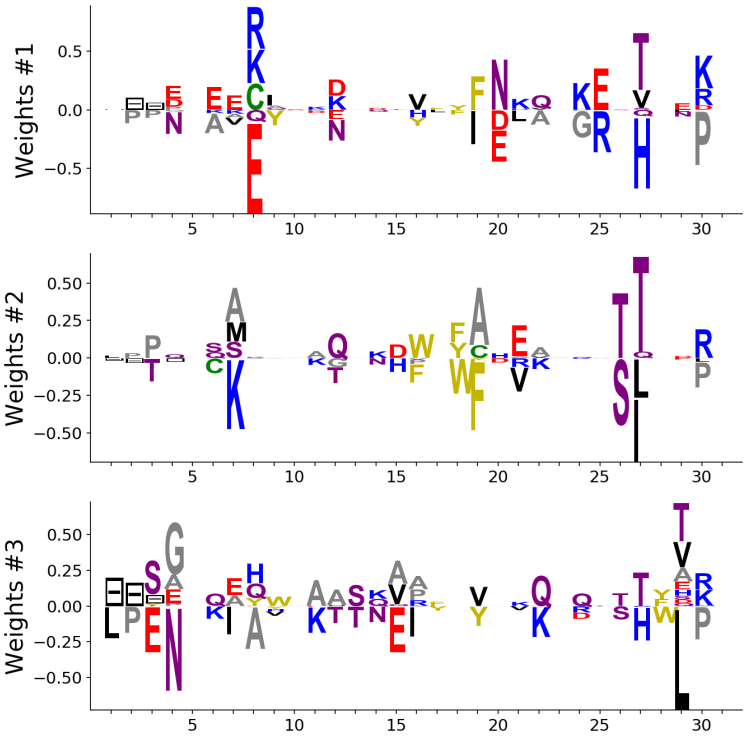


$\lambda_1^2 = 0$

$p = 0.68$

$\langle \mathcal{L} \rangle_{\text{Train}} = -0.86$

$\langle \mathcal{L} \rangle_{\text{Test}} = -1.49$



$\lambda_1^2 = 0.25$

$p = 0.14$

$\langle \mathcal{L} \rangle_{\text{Train}} = -1.24$

$\langle \mathcal{L} \rangle_{\text{Test}} = -1.35$

Increasing sparsity