Unsupervised learning of features from data: a statistical physics approach

R. Monasson
Laboratory of Physics
CNRS & Ecole Normale Supérieure, Paris

Model-guided data science, Como, September 2019

Plan of the lectures

1. Bayesian Inference and dimensional reduction:
   phase transition in principal component analysis

2. Representations: auto-encoders, Restricted Boltzmann
   Machines & sparse feature learning

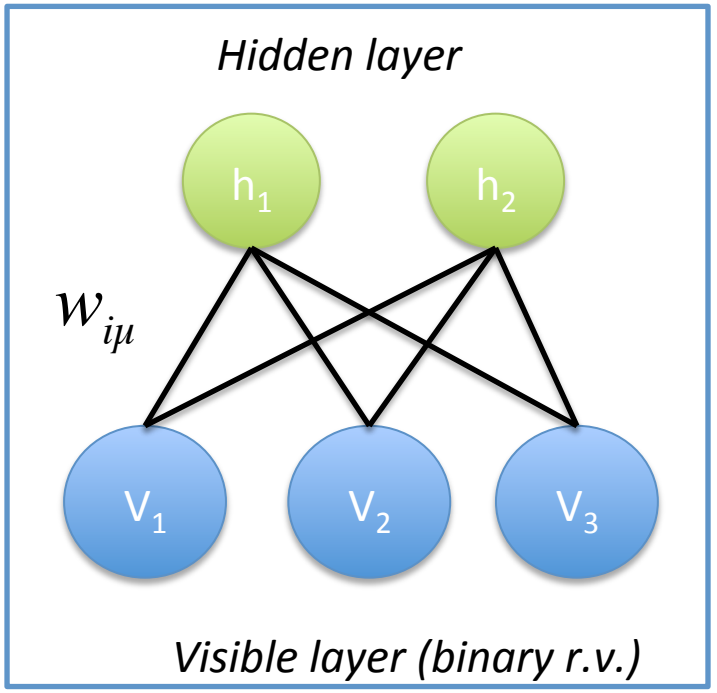
3. **Restricted Boltzmann Machines: connections with graphical models, phase transitions & applications**

Restricted Boltzmann Machines

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

$$P(v,h) = \frac{1}{Z}\exp\left[-E(v,h)\right]$$

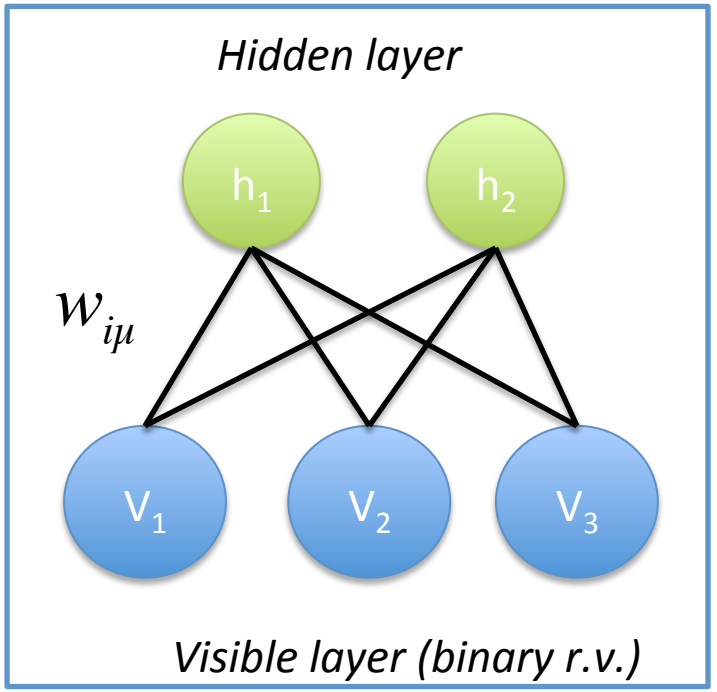
$$E(v,h) = -\sum_i g_i v_i + \sum_\mu U_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu$$



Smolensky 1986

Restricted Boltzmann Machines: Learning

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



$$P(v,h) = \frac{1}{Z}\exp\left[-E(v,h)\right]$$

$$E(v,h) = -\sum_i g_i v_i + \sum_\mu U_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu$$

- RBM learns a **probability distribution** over the **visible layer**.

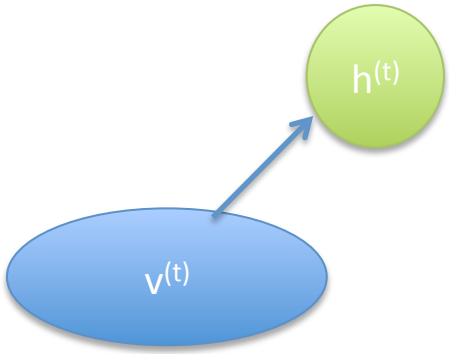
$$P(v) = \int \prod_\mu dh_\mu P\left(v,\{h_\mu\}\right) \equiv \frac{1}{Z_{eff}}\exp\left[-E_{eff}(v)\right]$$

- We also have
  - $P(h \mid v)$ extract (distribution of) representations from data
  - $P(v \mid h)$ generate (distribution of) data given a representation

RBM are generative models, trained through unsupervised learning

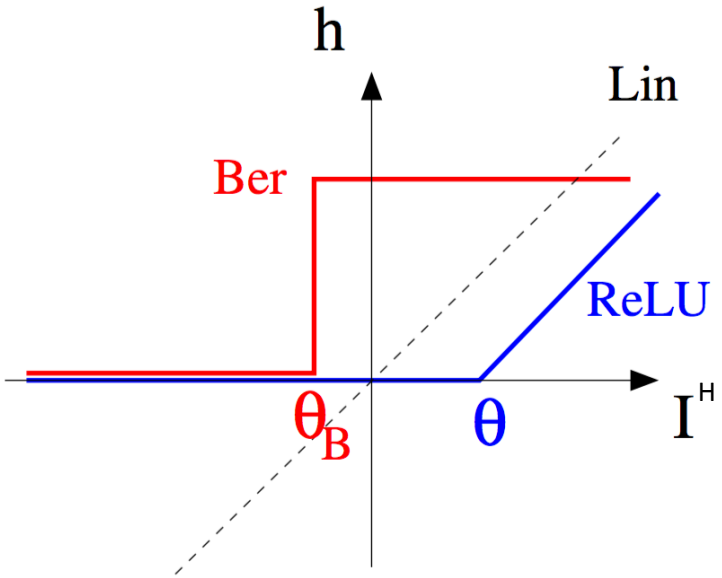
Hidden-Unit potential

- Compute hidden units inputs $I_\mu^H = \sum_i w_{i\mu} v_i$
- Sample hidden units $P(h_\mu \left| I_\mu^H \right.) \propto \exp\left[-U_\mu(h_\mu) + h_\mu I_\mu^H \right]$



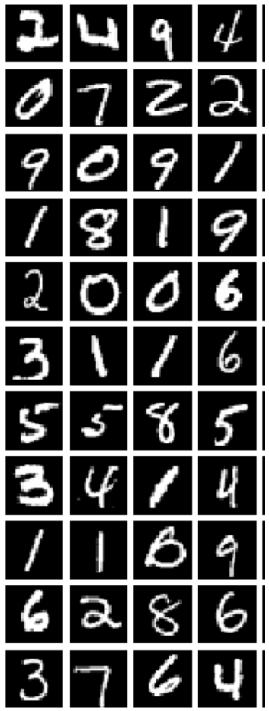
Most likely value of h given input I^H ?

$$\frac{d}{dh} U(h) = I^H \Rightarrow h = \Phi\left(I^H\right)$$



Bernoulli or ReLU empirically known to give better results than linear hidden units …

MNIST synthetic digits: Linear vs. ReLU RBMs



MNIST 60,000 image of digits of size 28x28

Linear RBM
M = 400
LogLikelihood : -0.15 bits/ pixel

ReLU RBM
M = 400
LogLikelihood : -0.11 bits/pixel

(digits are less noisy, more accurate, and MCMC are mixing faster)

RBM vs. BM: quadratic hidden potentials

Quadratic potential for hidden units: $U(h) = \frac{h^2}{2}$

1. Energy of visible and hidden configurations:

$$E(v,h) = -\sum_i g_i v_i + \sum_\mu U_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu} v_i h_\mu$$

2. Joint distribution:

$$P(v,h) = \frac{1}{Z}\exp\left[-E(v,h)\right]$$

3. Marginal distribution over visible configurations

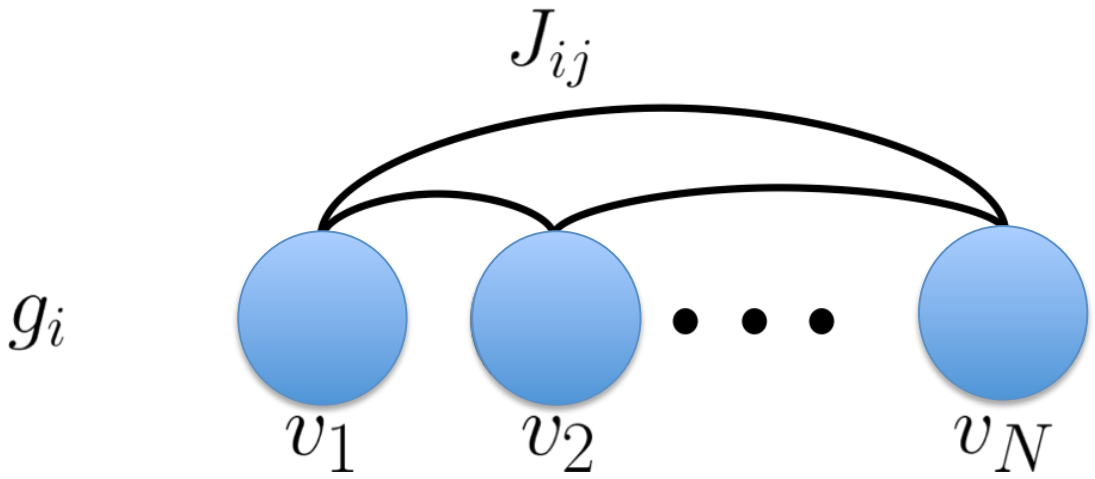
$$P(v) = \int \prod_\mu dh_\mu P\left(v,\{h_\mu\}\right) \equiv \frac{1}{Z_{eff}}\exp\left[-E_{eff}(v)\right]$$

RBM vs. BM: quadratic hidden potentials

$$E_{eff}(v) = -\sum_i g_i v_i + \frac{1}{2}\sum_\mu \left(\sum_i w_{i\mu} v_i \right)^2 = -\sum_i g_i v_i + \frac{1}{2}\sum_{i,j} \left(\sum_\mu w_{i\mu} w_{j\mu} \right) v_i v_j$$

J_{ij}

Boltzmann machine, i.e. Ising model with interaction matrix of rank-M !
Also called Hopfield model
Non quadratic potentials generate multi-body interactions between v_i ...



Talk by B. Bravi
next Friday

Hopfield's model (1982)

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554–2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. HOPFIELD

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, January 15, 1982

ABSTRACT Computational properties of use to biological organisms or to the construction of computers can emerge as collective properties of systems having a large number of simple equivalent components (or neurons). The physical meaning of content-addressable memory is described by an appropriate phase space flow of the state of a system. A model of such a system is given, based on aspects of neurobiology but readily adapted to integrated circuits. The collective properties of this model produce a content-addressable memory which correctly yields an entire memory from any subpart of sufficient size. The algorithm for the time evolution of the state of the system is based on asynchronous parallel processing. Additional emergent collective properties include some capacity for generalization, familiarity recognition, categorization, error correction, and time sequence retention. The collective properties are only weakly sensitive to details of the modeling or the failure of individual devices.

- Autoassociative memory
- Simple dynamics of components (no clock)
- Generalization, error correction, time sequence storage, …
- Robustness to failure of individual components

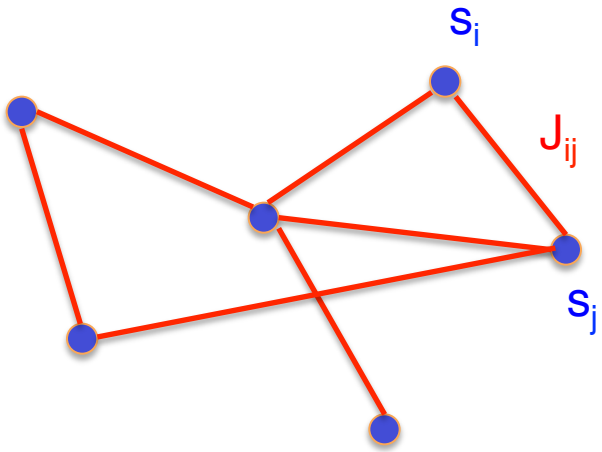
The model

- set of activity configurations (patterns) to be 'stored':

$$w_i^{\mu} = \pm 1$$

Index of pattern = 1, …, P

Index of neuron = 1, …, N



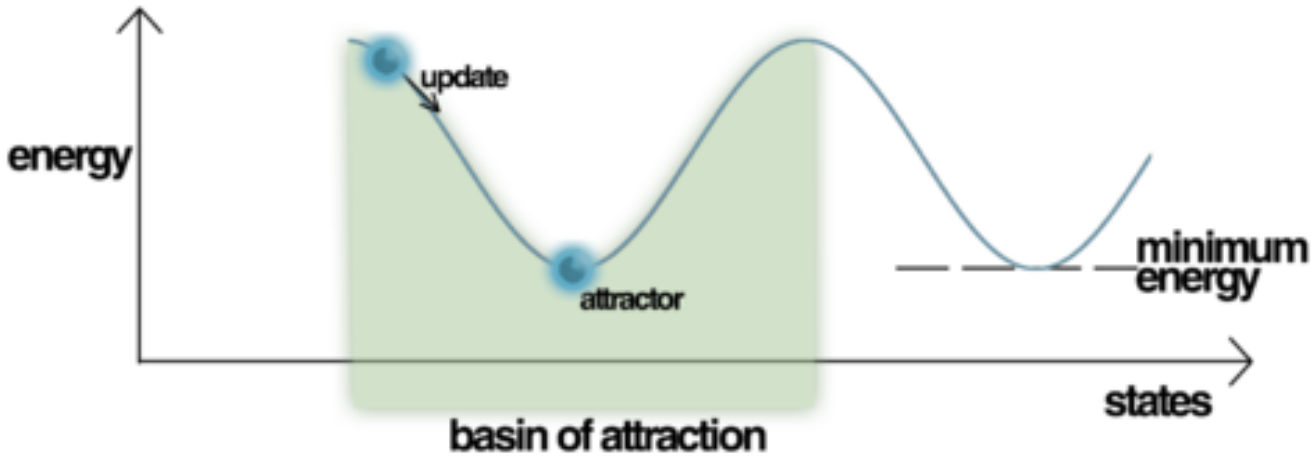
- synaptic interactions: $$J_{ij} = \frac{1}{N}\sum_{\mu} w_i^{\mu} w_j^{\mu}$$

- updating rule: $$v_i(t+1) = sign\left(\sum_j J_{ij} v_j(t) - \theta_i\right) \qquad (\theta_i = -g_i)$$

Q: Is the final state close to one of the patterns?

A: Yes, if number P of patterns small enough …

Dynamics is descending $E[v_1, v_2, \ldots, v_N] = -\frac{1}{2}\sum_{i,j} J_{ij} v_i v_j + \sum_i \theta_i v_i$



Are the minima of E close to the patterns?

Reminder on the mean field-Ising model

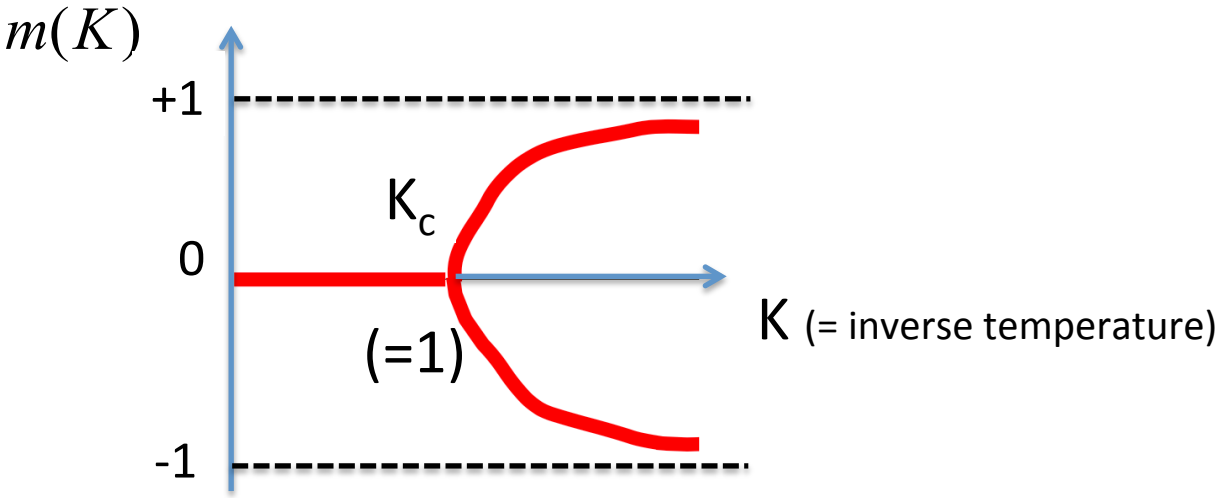
Probability of a configuration of spins:

$$P_K\left(v_1, v_2, \ldots, v_N\right) = \frac{1}{Z(K)} \exp\left(\frac{K}{2N} \sum_{i \neq j} v_i v_j\right)$$

Order parameter:

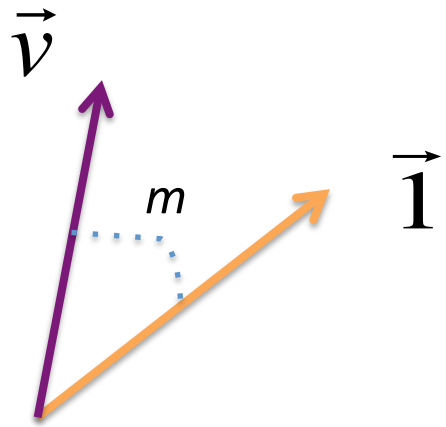
$$m(K) = \sum_{\{v_1, v_2, \ldots, v_N = \pm 1\}} P\left(v_1, v_2, \ldots, v_N\right) \times \frac{1}{N} \sum_i v_i$$

characterizes the degree of order of a typical configuration:

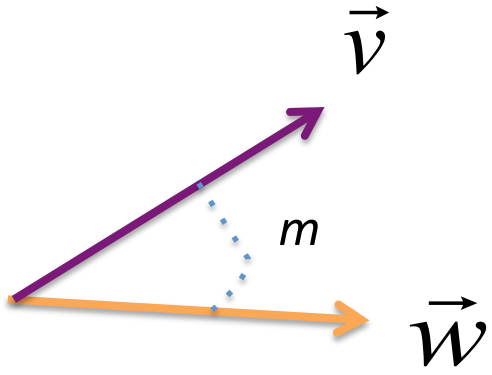


From the Mattis model…

$$P_{Ising}\left(v_1, v_2, \ldots, v_N\right) \propto \exp\left[\frac{K}{2N}\left(\sum_i v_i\right)^2\right] = \exp\left[\frac{K}{2N}\left(\vec{v}\cdot\vec{1}\right)^2\right]$$



(vertices of the N-dimensional hypercube)

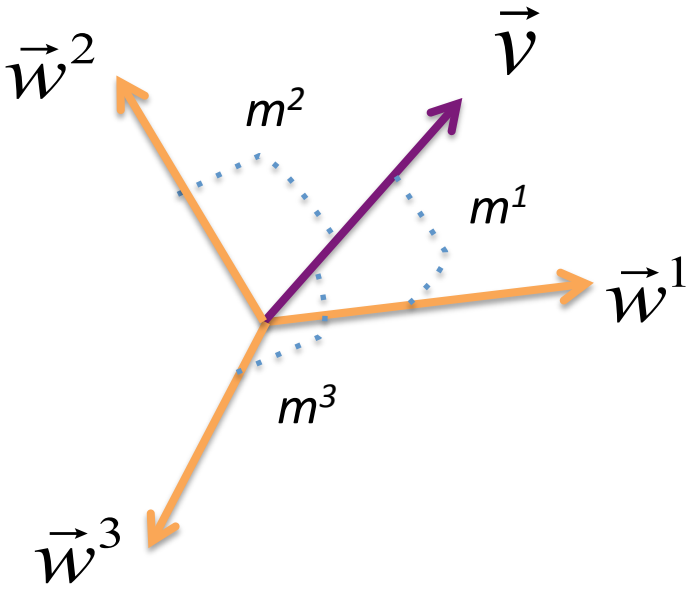


$$(w_i = \pm 1)$$

$$P_{Mattis}\left(v_1, v_2, \ldots, v_N\right) \propto \exp\left[\frac{K}{2N}\left(\sum_i v_i w_i\right)^2\right]$$

Same phase transition!

… to the Hopfield model



p « patterns »

$$\vec{w}^{\mu}$$

$$\left(\mu = 1 \ldots p\right)$$

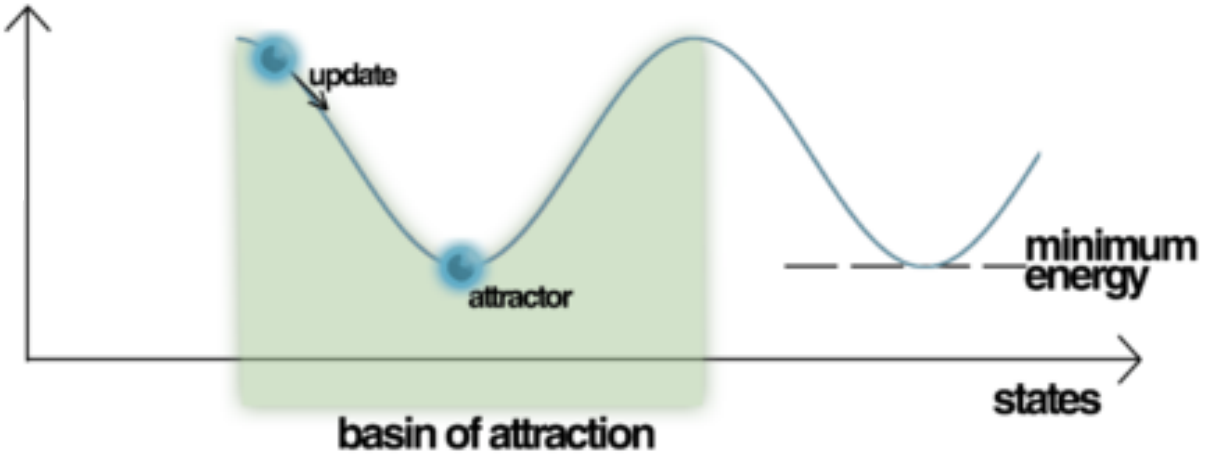
$$P_{Hopfield}\left(v_1, v_2, \ldots, v_N\right) = \frac{1}{Z\left(\left\{w_i^{\mu}\right\}\right)} \exp\left[\frac{K}{2N} \sum_{\mu=1\ldots p}\left(\sum_i v_i w_i^{\mu}\right)^2\right]$$

Same phase transition as Ising model if p finite & N tends to infinity:

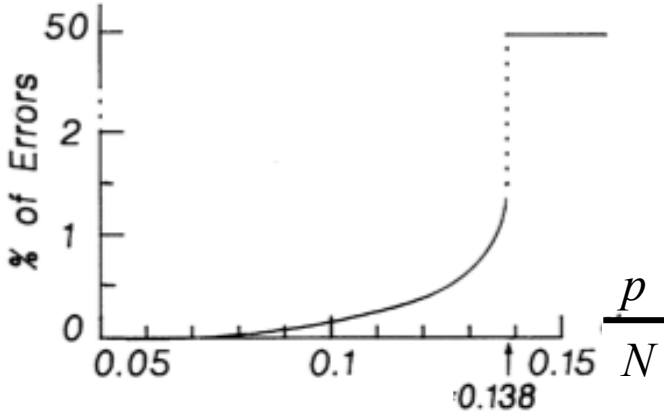
i.e., for large *K*, one overlap m^{μ} is positive, the others vanish (other scenarios with multiple positive overlaps are exponentially unlikely…)

The Hopfield model in the double p,N ➔∞ limit

$$E\left[v_1, v_2, \ldots, v_N\right] = -\frac{K}{2}\sum_{i,j}\left(\sum_{\mu} w_i^{\mu} w_j^{\mu}\right) v_i v_j$$



Amit, Gutfreund, Sompolinsky (1985)

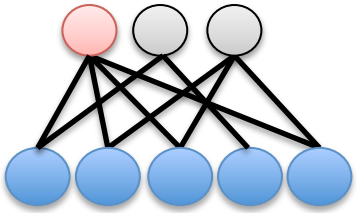


$$\left.\frac{p}{N}\right|_c \approx 0.138$$

(large K)

Gaussian RBM: The Phases (at low temperature)

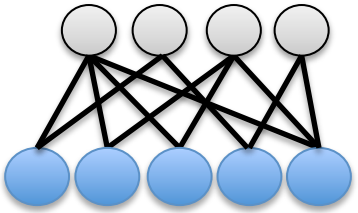
Ferromagnetic Phase :



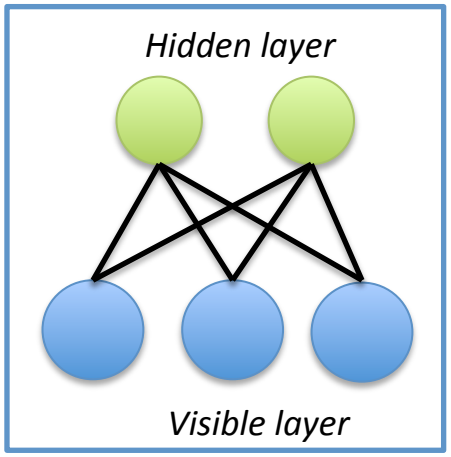
- One hidden unit is strongly activated and the others are weakly activated.
- Number of regions with high probability is linear in N

More hidden units

Glassy Phase :

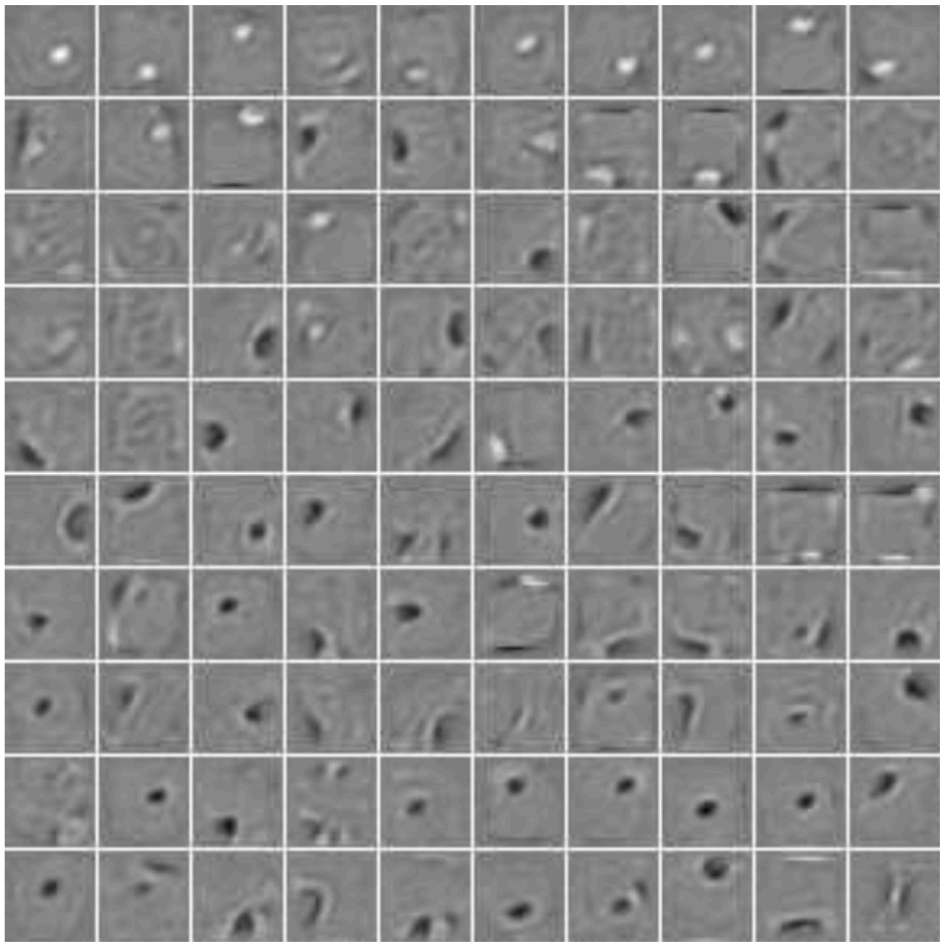


- All hidden units are weakly activated
- All states have weak probability



Prototype vs. compositional regimes in nonlinear RBM

16 hidden units (**prototypes)**

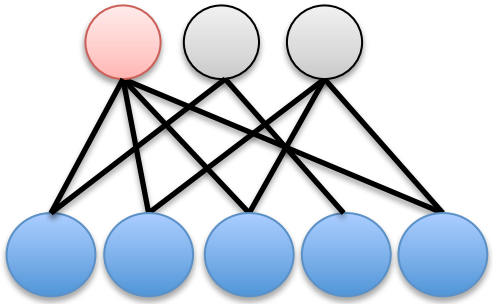


100 hidden units (**sparse features**)

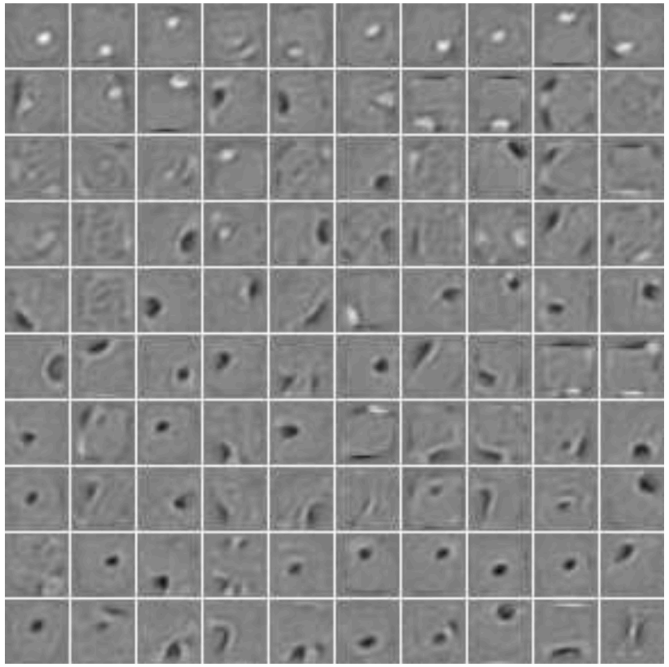
Fischer & Igel. Training Restricted Boltzmann Machines: An Introduction, 2014.

Prototype regime

Strongly reminiscent of ferromagnetic regime in Hopfield model :



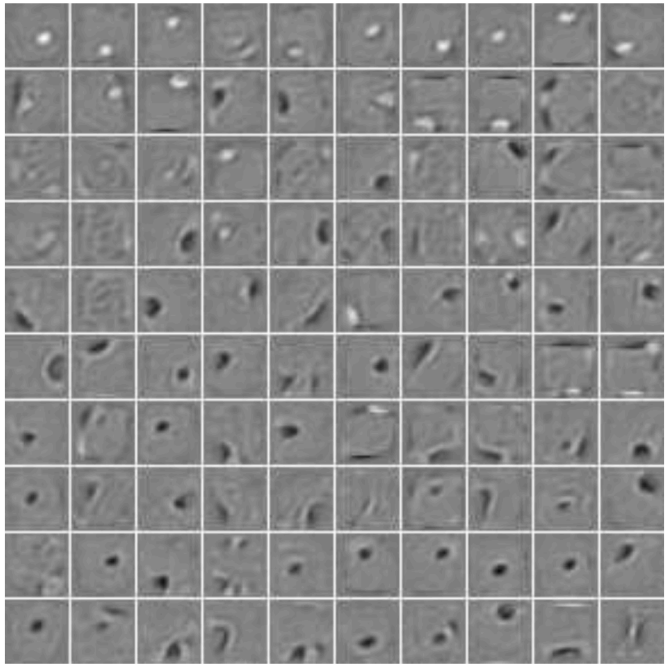
- One hidden unit is strongly activated and the others are weakly activated;
- Limited diversity of visible-layer configurations given hidden-unit activation
- Number of high-probability configurations is linear in nb. of hidden units

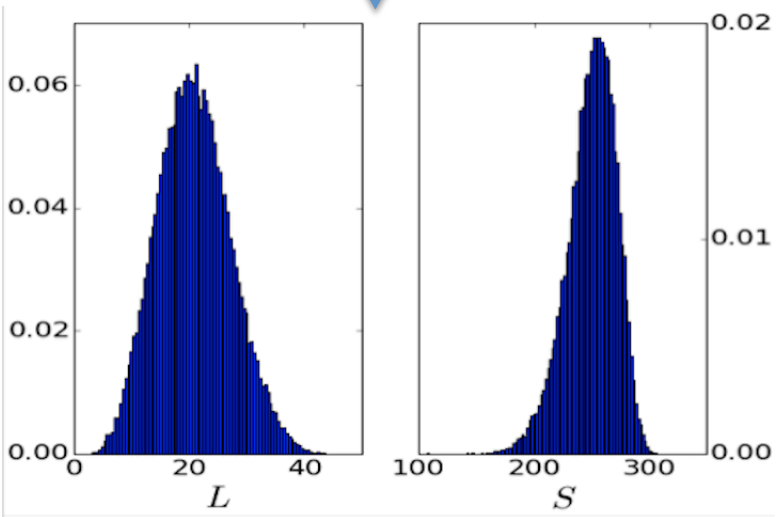


Compositional regime

??

Compositional regime



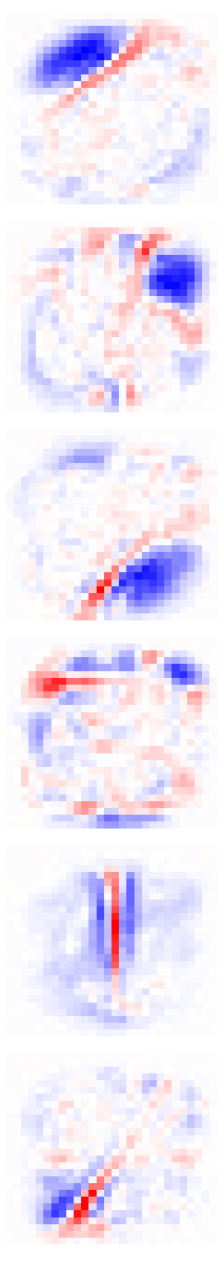


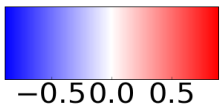
Each generated digit image is composed by superposing about L≈20 elementary strokes, while S≈250 units are silent. Different combinations of strokes produce different variants of the same digit or different digits

A « dictionary » of sparse and strong features

A subset of the features learnt for a ReLU RBM. M = 400.

Each image represent a weight vector $\{w_i\}_\mu$





- After learning of data, weights are sparse
- And few, remaining weights are large!

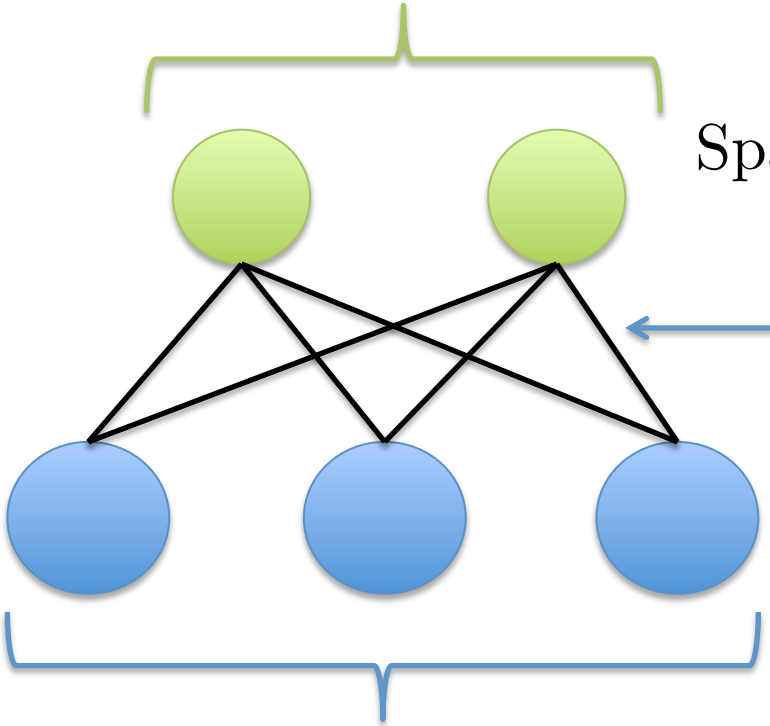
Conditional averages Black: v=0, White: v=1, Grey scale

- No interference thanks to large ReLU threshold …

Random-Weight RBM ensemble

$h_\mu \in \mathbb{R}_+$, ReLU+
Threshold ϑ

αN



Sparse Random Weights

$$W \begin{cases} 0 & 1-p \\ +1 & \frac{p}{2} \\ -1 & \frac{p}{2} \end{cases}$$

$v_i \in \{0,1\}$
field g

$N \to \infty$

Statistical Mechanics of Random-Weight RBMs

- Replica theory computation is performed to estimate the average free energy F in the zero-temperature limit :

$$F(\alpha, p, g, \theta) = \lim_{\substack{N\to\infty \\ T\to 0}} -\frac{T}{N} \log \sum_{\{v_i\}} \int \prod_{\mu} dh_\mu \exp\left[-E(v,h;W,g,\theta)/T\right]$$

Number of hidden units / Number of visible units

Fields on visible units

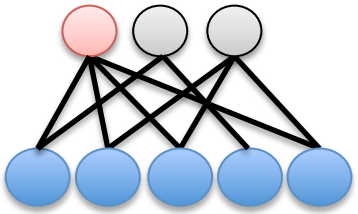
Weight sparsity

Hidden unit threshold

Averaged out

The three representational regimes of RBM

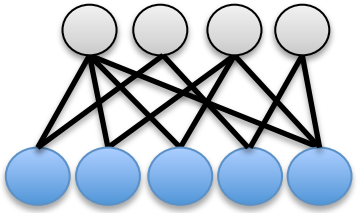
"Ferromagnetic" regime



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

More hidden units

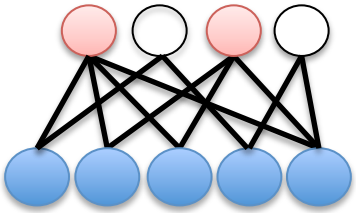
"Glassy" regime



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

Increasing sparsity

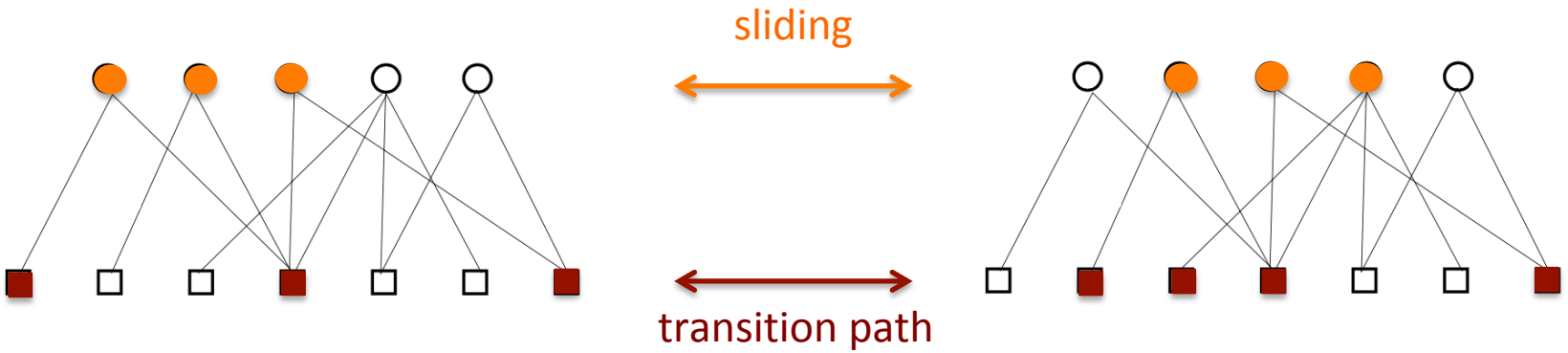
Non quadratic hidden-unit potentials



Compositional regime

- Multiple hidden units very active
- Corresponding weights define features composing visible configurations
- Extract invariances; interpretable

Transition paths



Example of Transition Path

(diffusive motion)

Proteins: structure & function

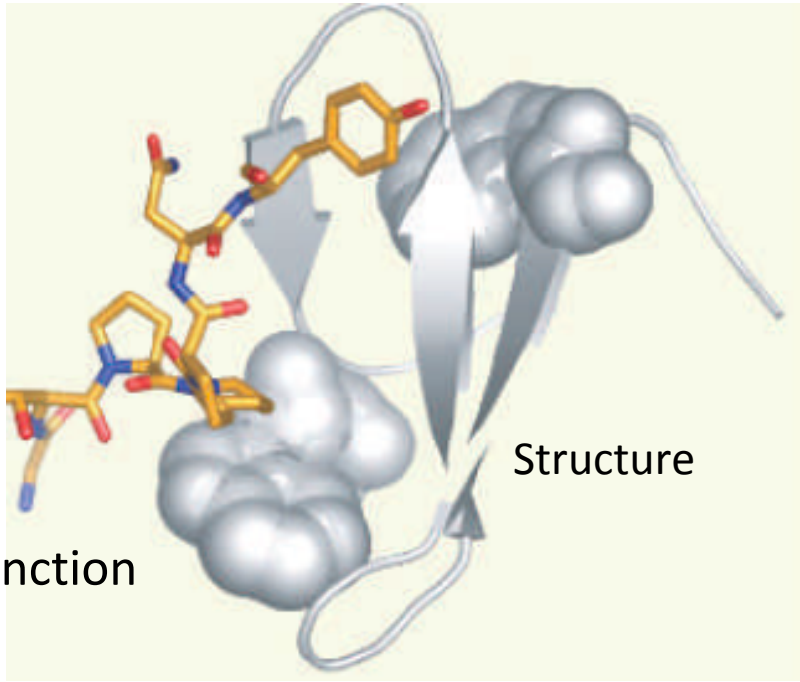
(with S. Cocco, J. Tubiana)

Amino-acid Sequence

ALPPGWEERIHLDGRTFYIDHNSKITQWEDPRLQ

Function

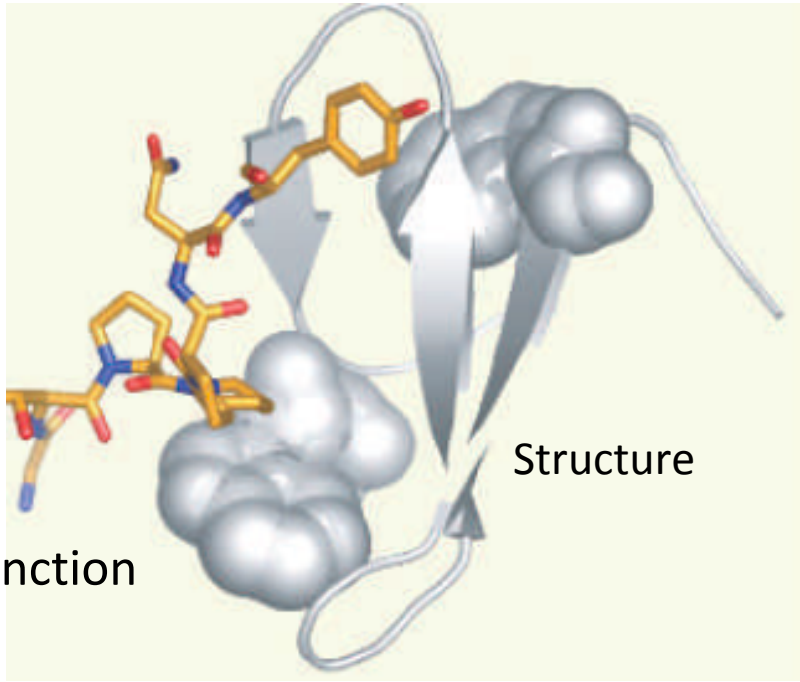
WW protein domain



- Sequence fully defines 3D structure and function of the corresponding protein: Genotype-phenotype mapping?

Proteins: structure & function

WW protein domain



Amino-acid Sequence

ALPPGWEERIH**L**DGRTFYIDHNSKITQWEDPRLQ

Mutation: L ➔ A

- **neutral**: no effect
- **detrimental**
- **beneficial**

?

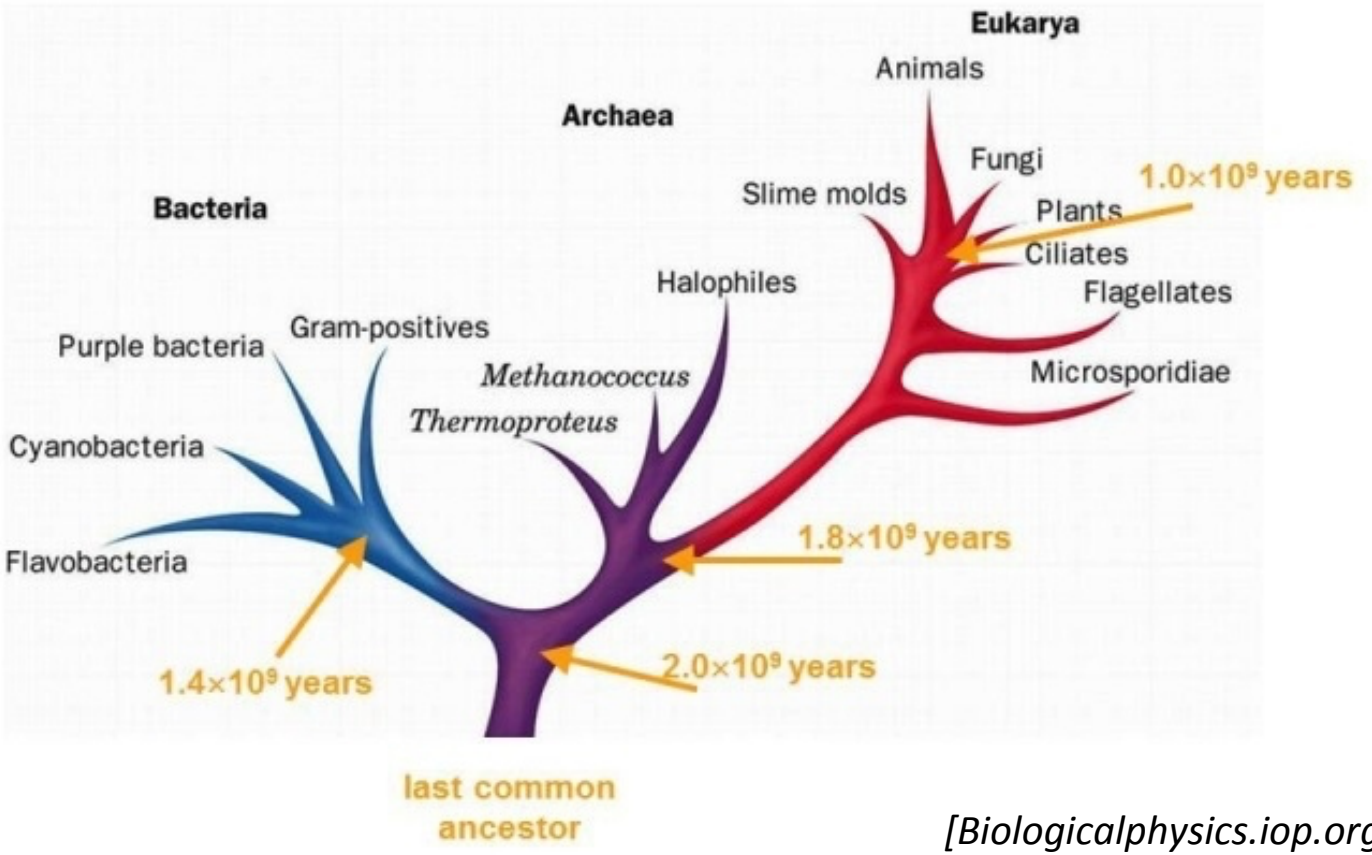
Function

Structure

- Sequence fully defines 3D structure and function of the corresponding protein: Genotype-phenotype mapping?

- Mapping is complex:
  Mutations may largely affect stability, activity, specificity, … of protein
  (directed evolution for protein design)
  No additivity a priori: epistasis

Proteins: Evolutionary Families



[Biologicalphysics.iop.org]

- Many realizations of sequences of proteins with similar function & structure
- Characterization of underlying distribution??
- May help to Infer structure, function; Predict effect of mutations; Design new proteins; …

Multi-sequence alignments

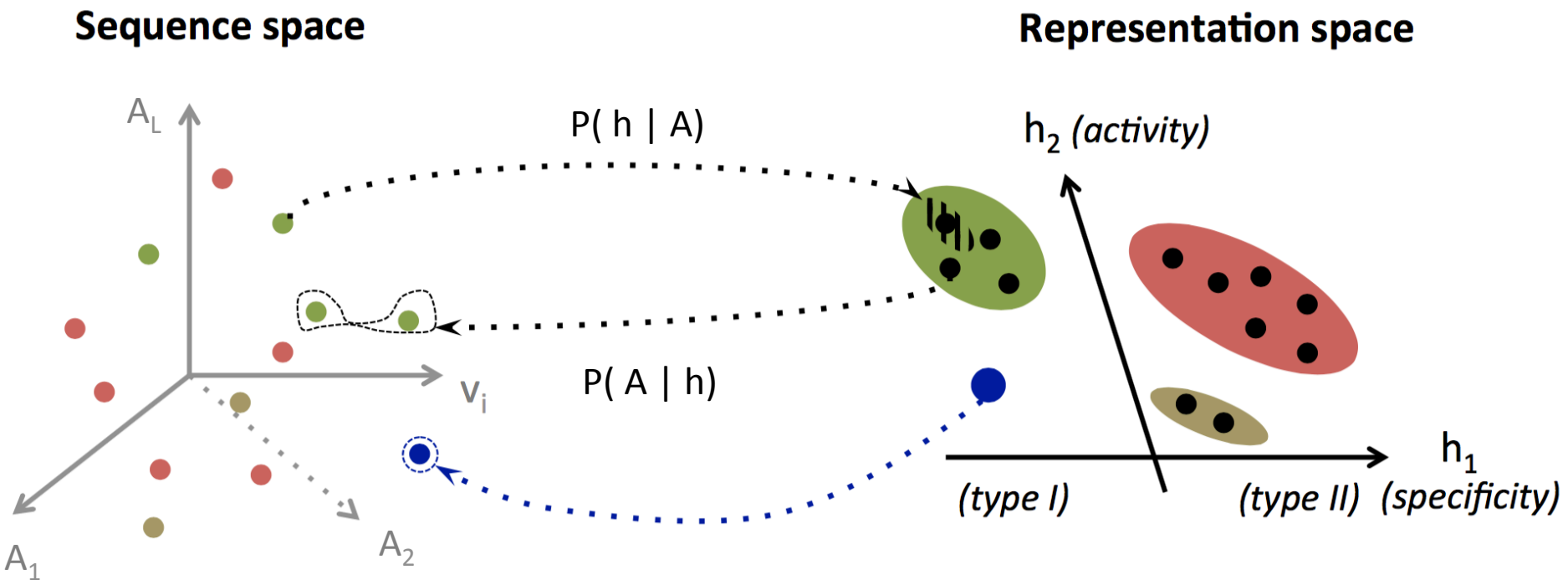
```
Q9U3M7_CAEEL/420-758  YQ-------- ---------- -YPRYVTE-I TVRHLLEHTA GGWDNLQSDA
 O62481_CAEEL/42-406  YLPEF----- --------PA KKFKNEDVKI TMRQLLSHSA GIRHYATEKK
 Q9I1G0_PSEAE/16-386  WLPAF----- --------RP RLADGREARI TPRQLLSHSA GLGYRFLEAD
   ESTB_BURGA/22-391  WLPEF----- --------RP RLADGSEPLV TIHHLLTHTS GLGYW-LLEG
   LOVD_ASPTE/17-406  RLLPDLSAMP VLEGFDDAGN ARLRERRGKI TLRHLLTHTS GLSYVFLHPL
 Q9WXD6_BRELN/11-387  YAPGLADVQV I-EGFDVDGS PILRAPASEP TTKQLLLHTA GFGYDFFNEK
 Q9L8D3_SORCE/35-419  WLPELANRKV --LARIDGPI DETVPAERPI TVRDLMTFTM GFGISFDASS
 Q9A2D7_CAUCR/97-472  YLPEFADM-- ----KVSDGQ GGVRPAARPI LVRDILRHTA GFSYGWGEGP
 Q9A800_CAUCR/77-466  HIPEFANLRV A-KGVDETGQ PILTPVSRSP TMRELMTHTA GFAYGLANDP
 Q9A7Z9_CAUCR/50-429  FIPEFANLRV L-KGVNADGS FDTVPAERPP TMRELMSHSA GFAYGLTPDN
 O88012_STRCO/23-405  YLPAFAEPRV --YVGGTGEN VVTRPATGPV RVRHLMTHTA GLTFGFYRTH
 Q9A3L3_CAUCR/99-483  FIPAWRDIGV --FQAGVAGA FQTTRTKEPM RTIDLLRHTS GLTYGFQQRT
 P72041_MYCTU/40-404  YLP------- ---------- SYTSHGKHRT TIRHVLTHSA GVPFPTGPRP
 Q9X5Q0_STRLA/21-361  YWP------- ---------- QFARHGKGDV TVRHVLQHRA GVPVGRGIVR
 Q18384_CAEEL/50-416  YWP------- ---------- EFGQNGKQDI TIEMVLSHTA GLPYFPGVKF
 Q9XU43_CAEEL/52-422  YWP------- ---------- EYGRYGKNAT TIEDVLSHKA GLPYL-SEDI
```

Sequence number

Amino-acid content

[- = gap, same format for all sequences]

High-dimensional representations of protein sequences



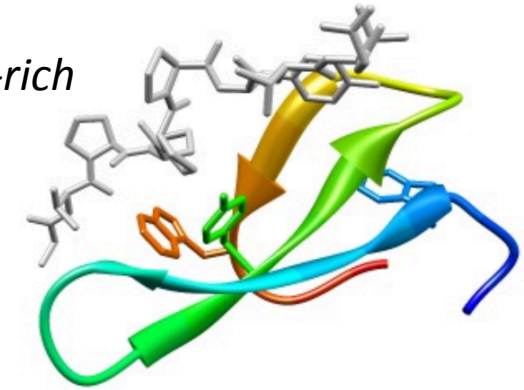
- RBM extract high-D representations of (common inputs to) sequences
- Representations are useful (to design « good » sequences) …
- … and, hopefully, biologically meaningful (structure, function, history)

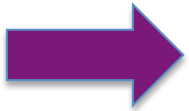
➔ Practical implementation of genotype-to-phenotype relation

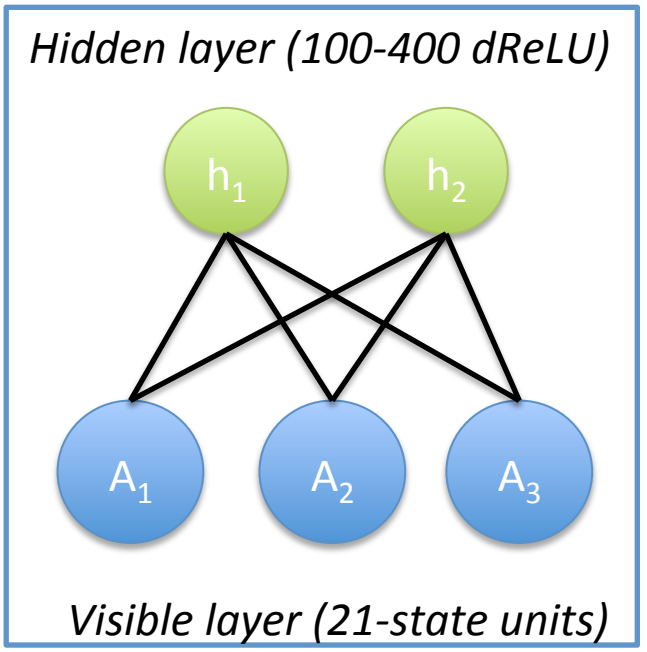
Applications of RBM to protein sequence data

WW domain (PFAM PF00397)

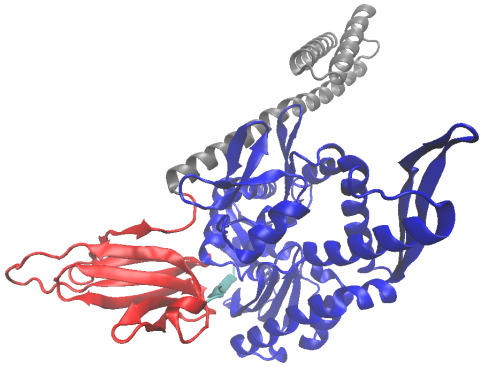
Proline-rich ligand



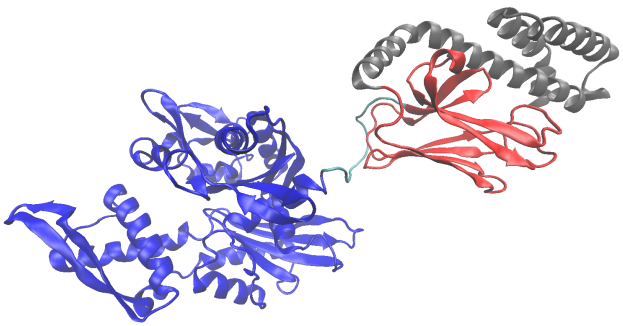




HSP70 chaperone protein (PFAM PF00012)



ATP-bound

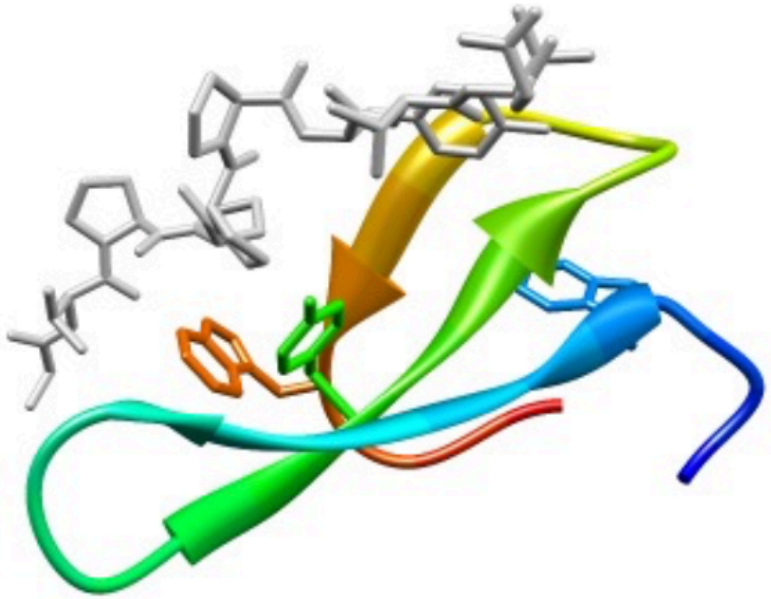


ADP-bound

Tubiana, Cocco, R.M., eLife 2019

Example 1: the WW domain

- Binding domain involved in eukaryotic signalling proteins
- N=30-40 amino-acids (very small)
- Folds into 3-stranded antiparallel beta strands



- Recognizes Proline-rich Linear Motifs with 4 types of ligand specificities

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

Example 2: HSP70 chaperone protein

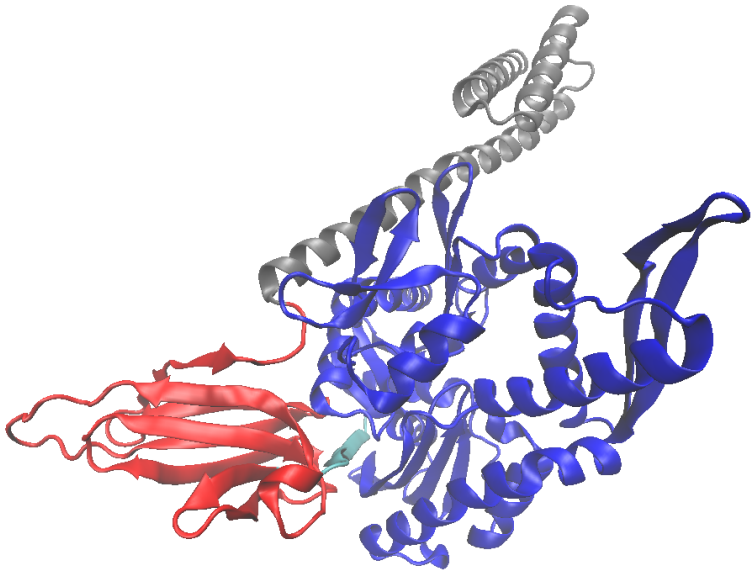
- N>600 amino-acids
- Multidomain.
  - Nucleotide Binding Domain (NBD)
  - Substrate Binding Domain (SBD)
  - LID Domain
  - Linker

Function:

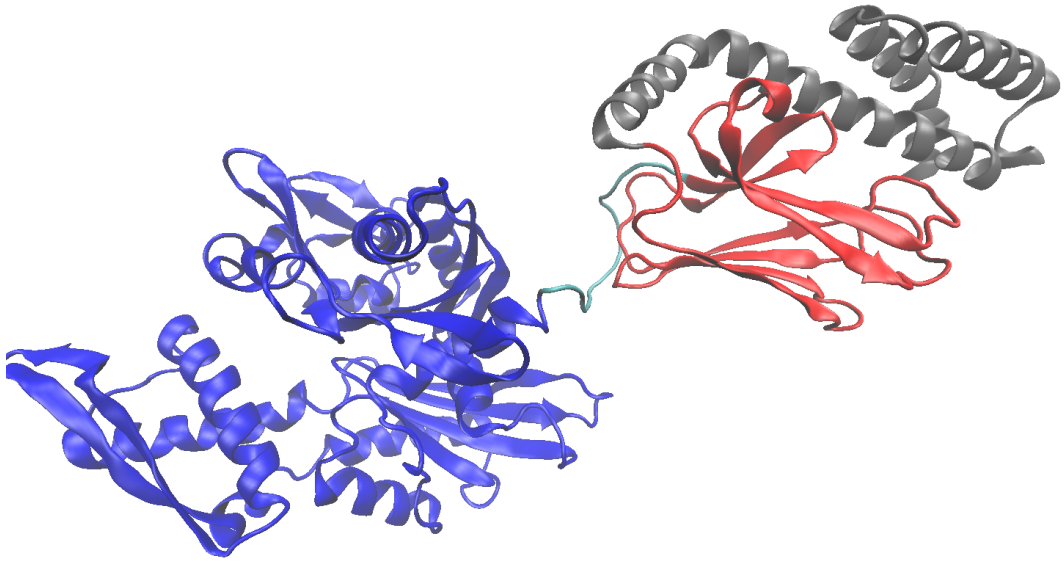
- Traps substrate proteins between the LID and the SBD
- LID/SBD cavity is either open or close

Roles:

- Assist protein folding
- Transport proteins for degradation

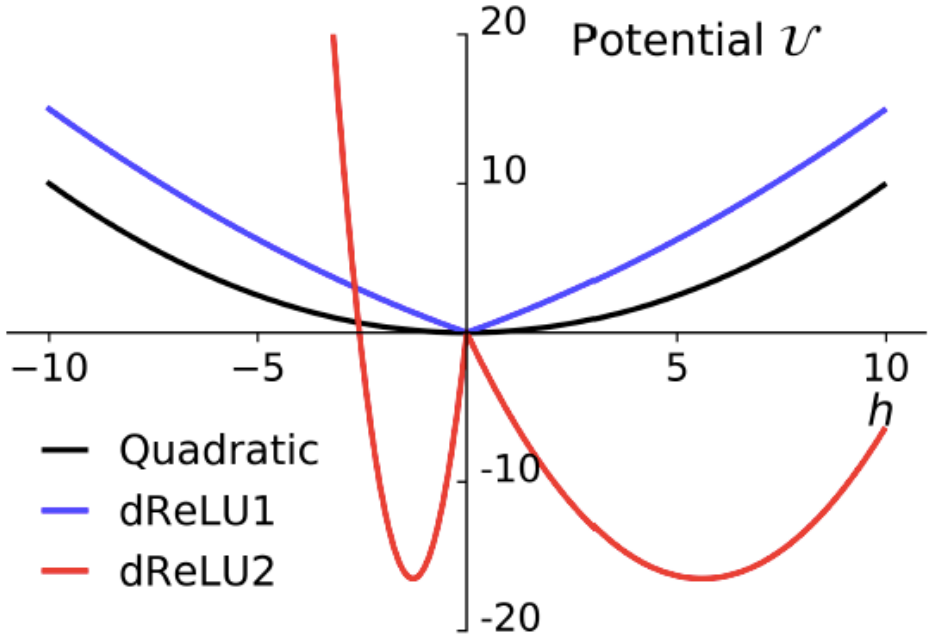


ATP bound conformation (open)
PDB: 2kho



ADP bound conformation (closed)
PDB: 4jne

Hidden-unit potentials

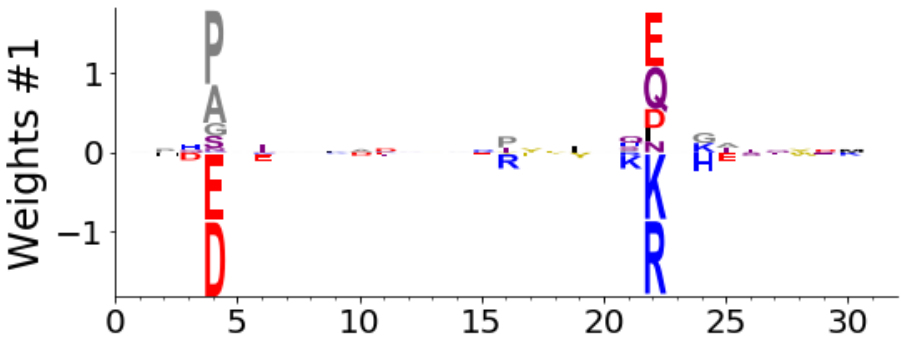


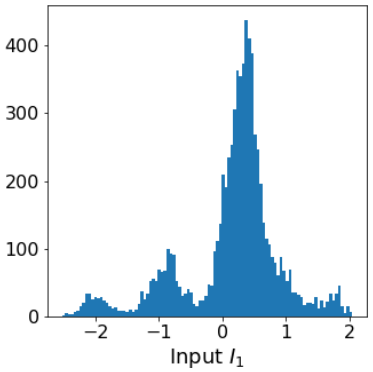
Generalization of Rectified Linear Units:

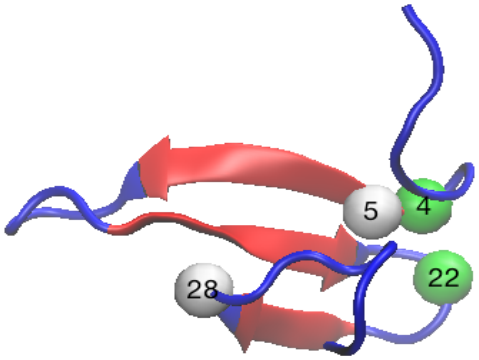
- Positive slopes at origin favors h=0 (as L1 regularization)
- Negative slopes favors bimodal distribution for h
- Confining potential at large h
- Four parameters to be learned from data for each hidden unit

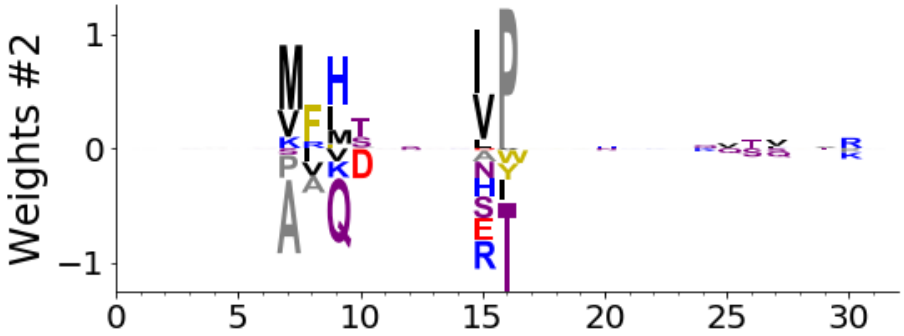
Weights reflect structural features

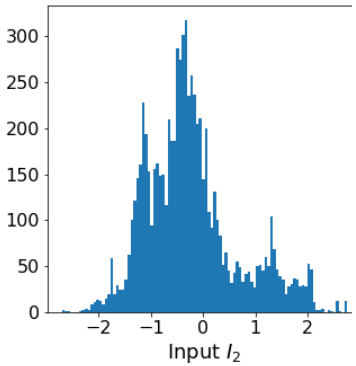
WW domain

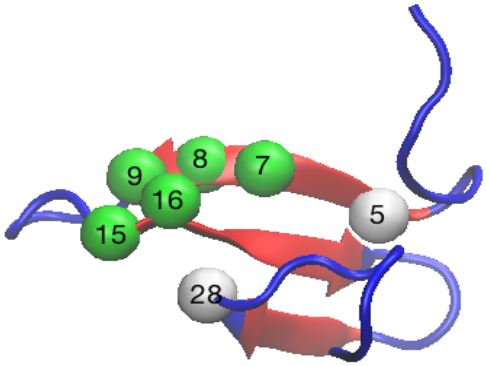






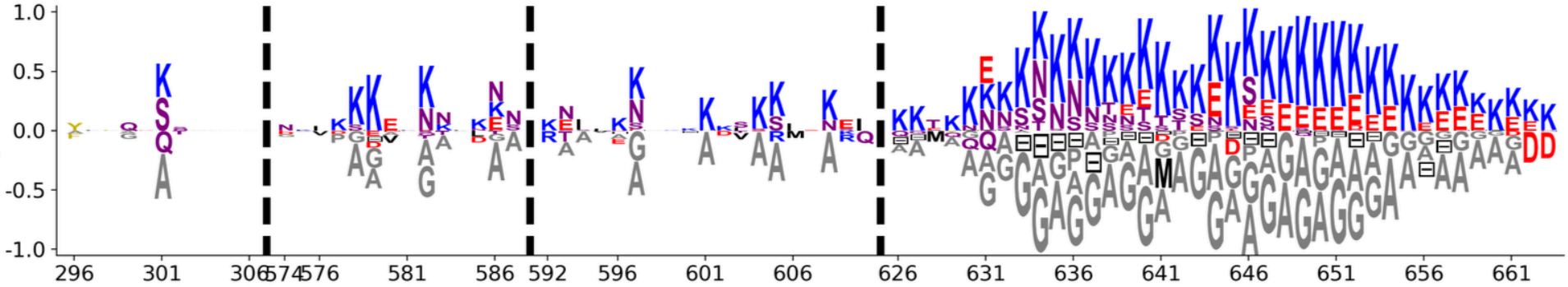






Weights reflect structural features

HSP70

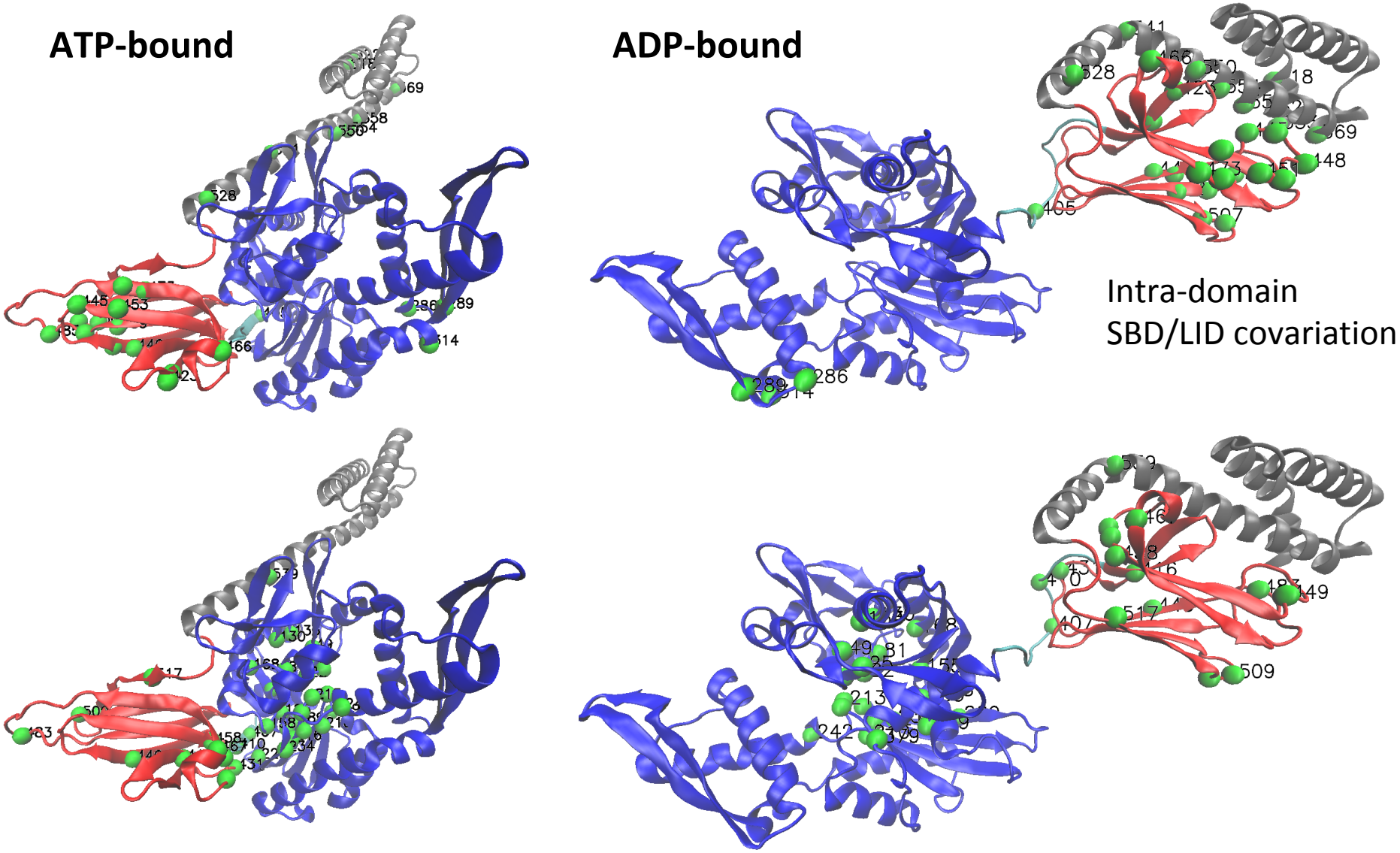


- Collective mode located on the unstructured C- terminal tail
- Known to be crucial for interactions with co-chaperones
- Either charged hydrophilic or tiny hydrophobic residues
- Analogous to IDP

Interdomain weights control allostery

ATP-bound

ADP-bound



Intra-domain
SBD/LID covariation

Nucleotide Binding Domain (NBD)

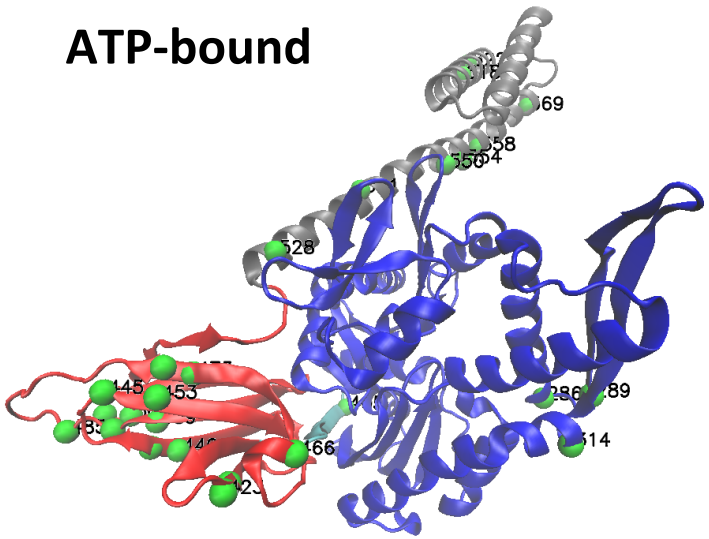
LID Domain

Substrate Binding Domain (SBD)

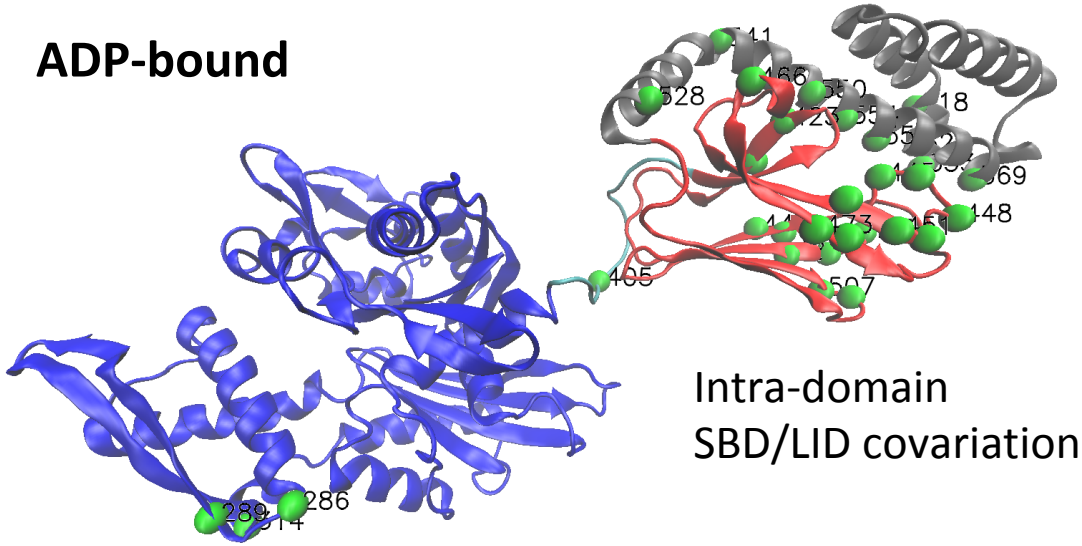
Linker

Interdomain features control allostery

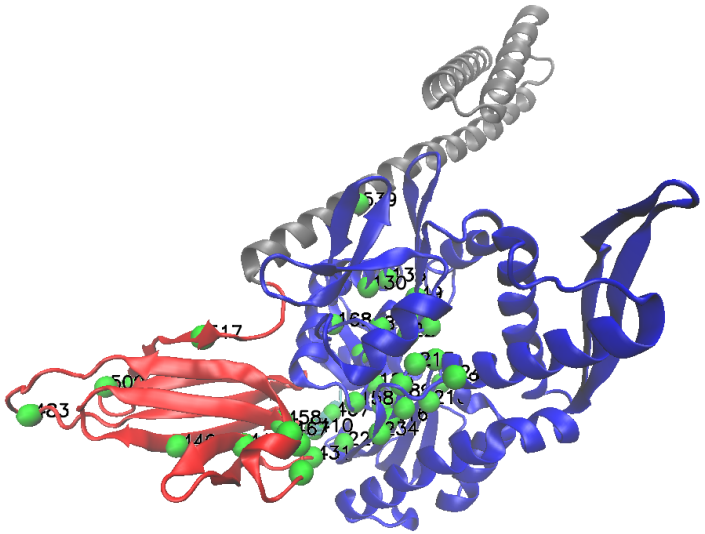
ATP-bound

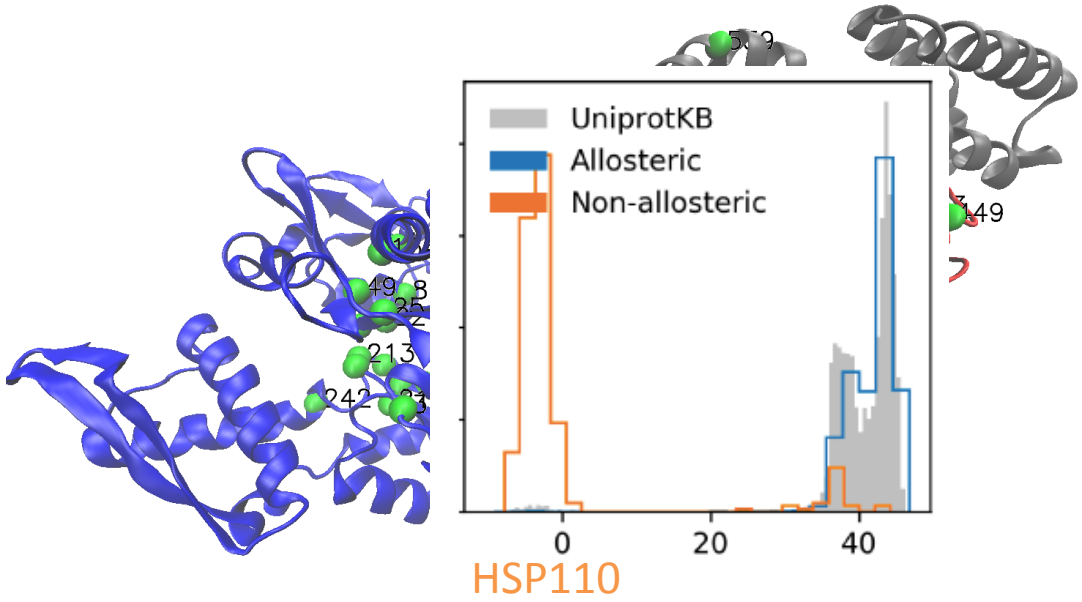


ADP-bound



Intra-domain
SBD/LID covariation

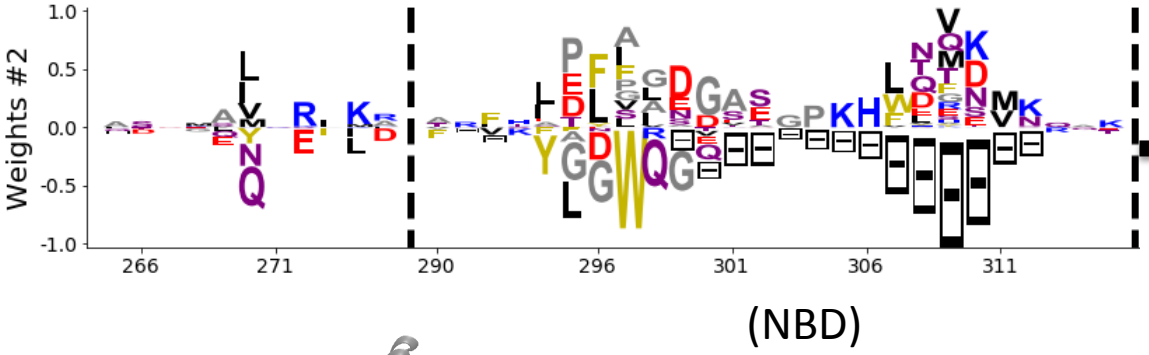




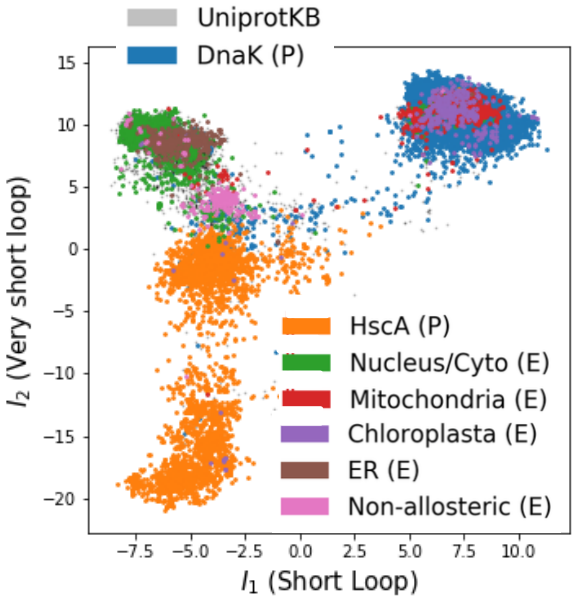
HSP110

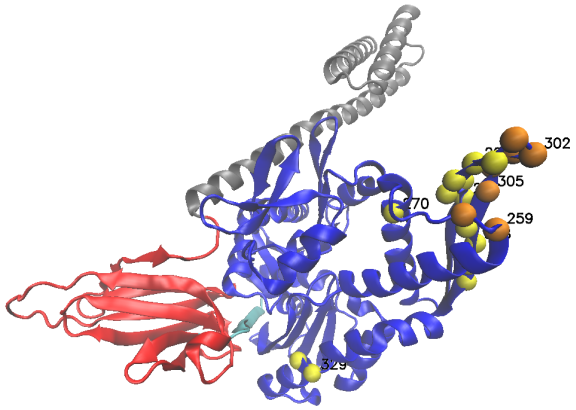
Non-allosteric specific

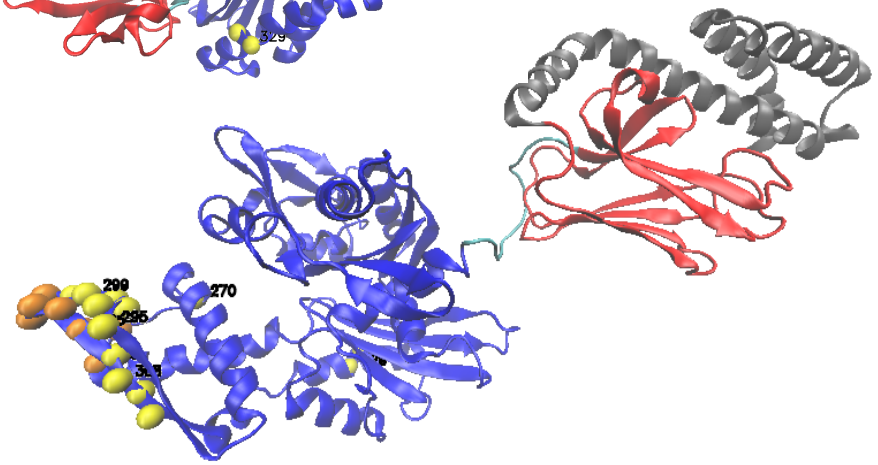
Loop motifs control ATP/ADP regulation

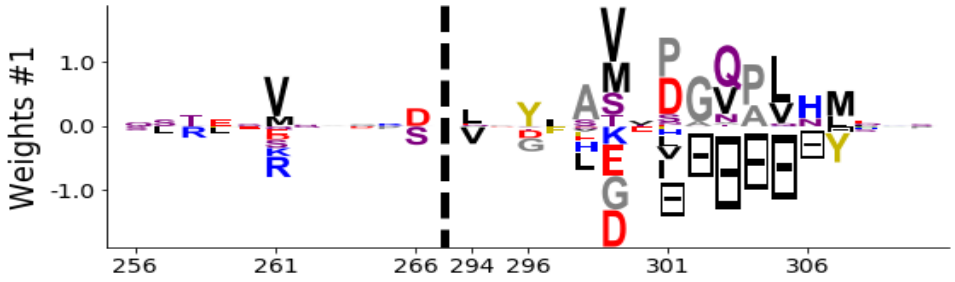


(NBD)

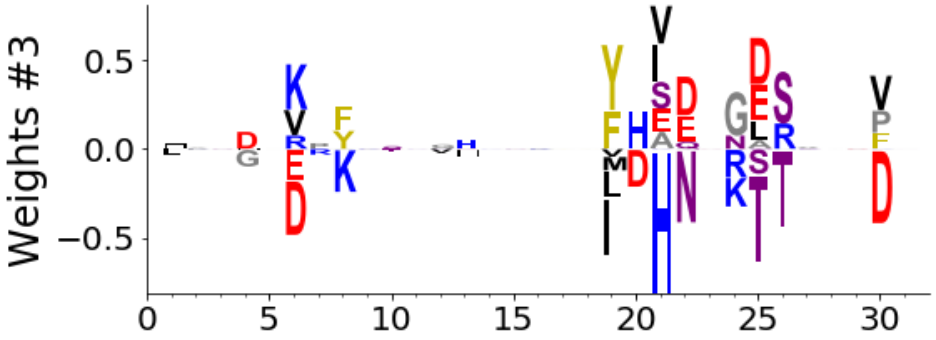


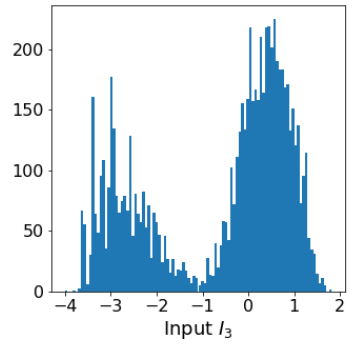


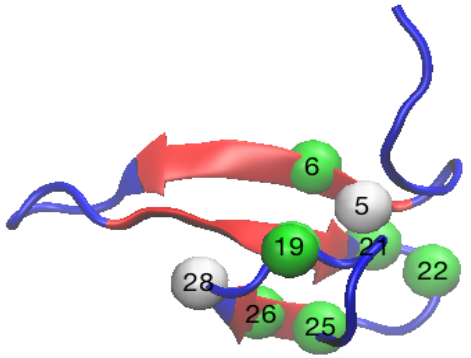


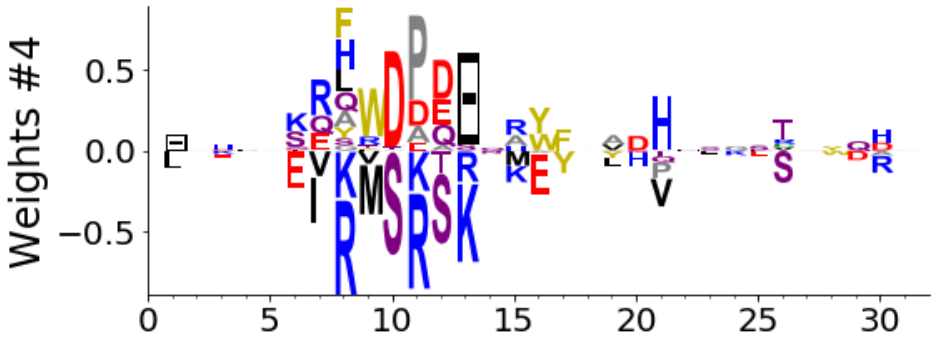


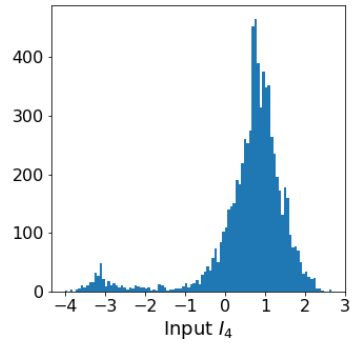
RBM features reflect function

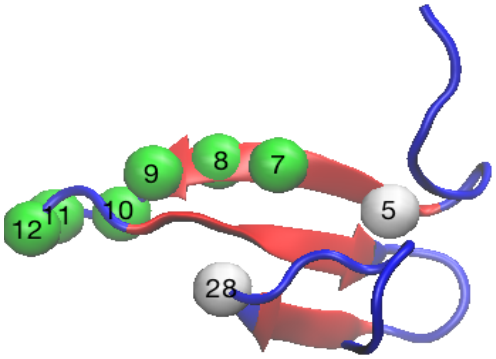








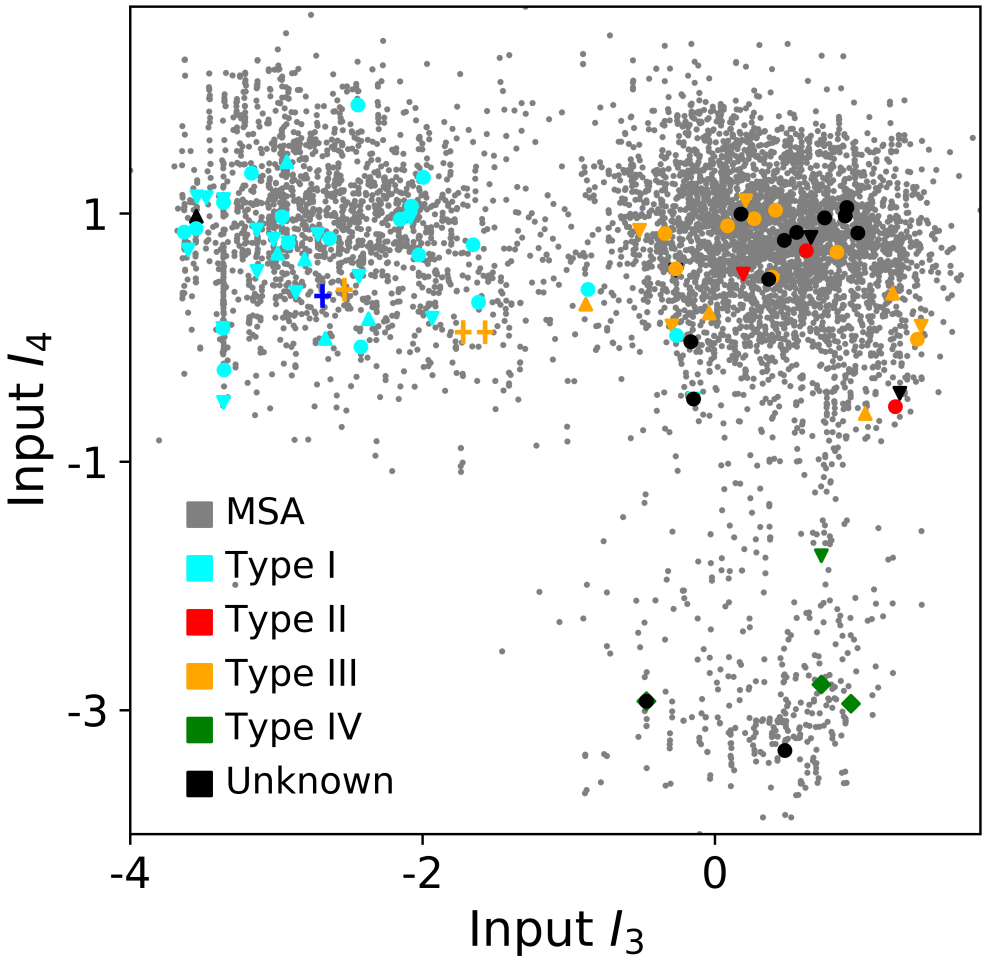




WW Domain: Weights Determine Motif Recognition Specificity

Motif recognized

Type I : PPXY
Type II : PPLP
Type III: PR
Type IV: [p(S/T)P]

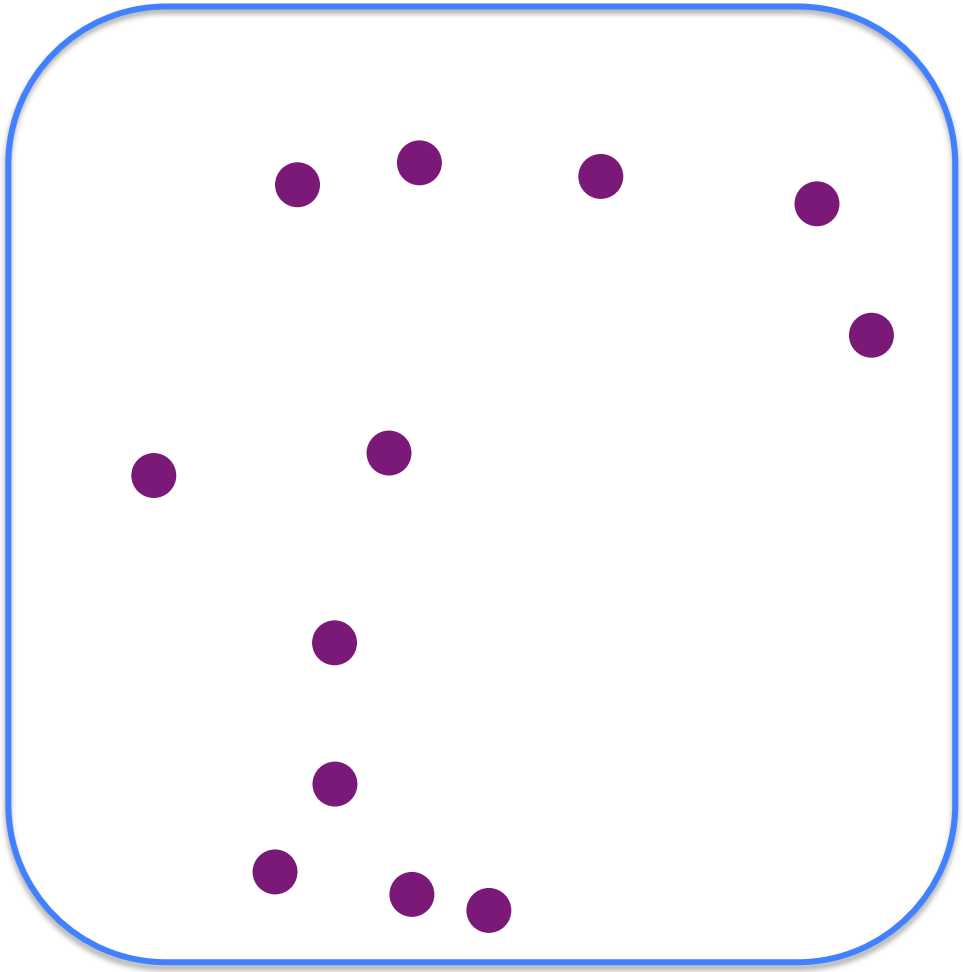


Experimental data from:
Russ et al. Nature 2005
Espanel and Sudol J. Biol. Chem. 1999
Otte et al. Protein Science 2003

Why are RBM extracting useful/interpretable representations?

The problem:

Find probability distribution from very few samples

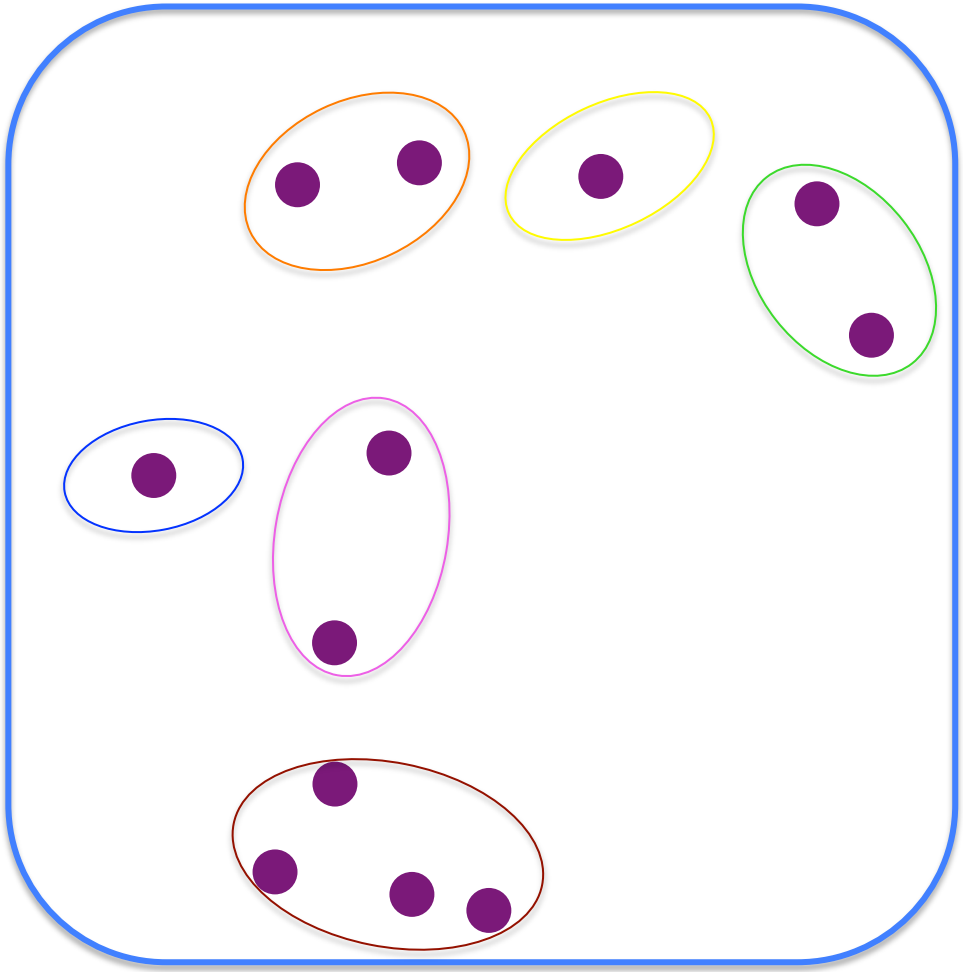


Sequence space

Why are RBM extracting useful/interpretable representations?

Mixture of local models :

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



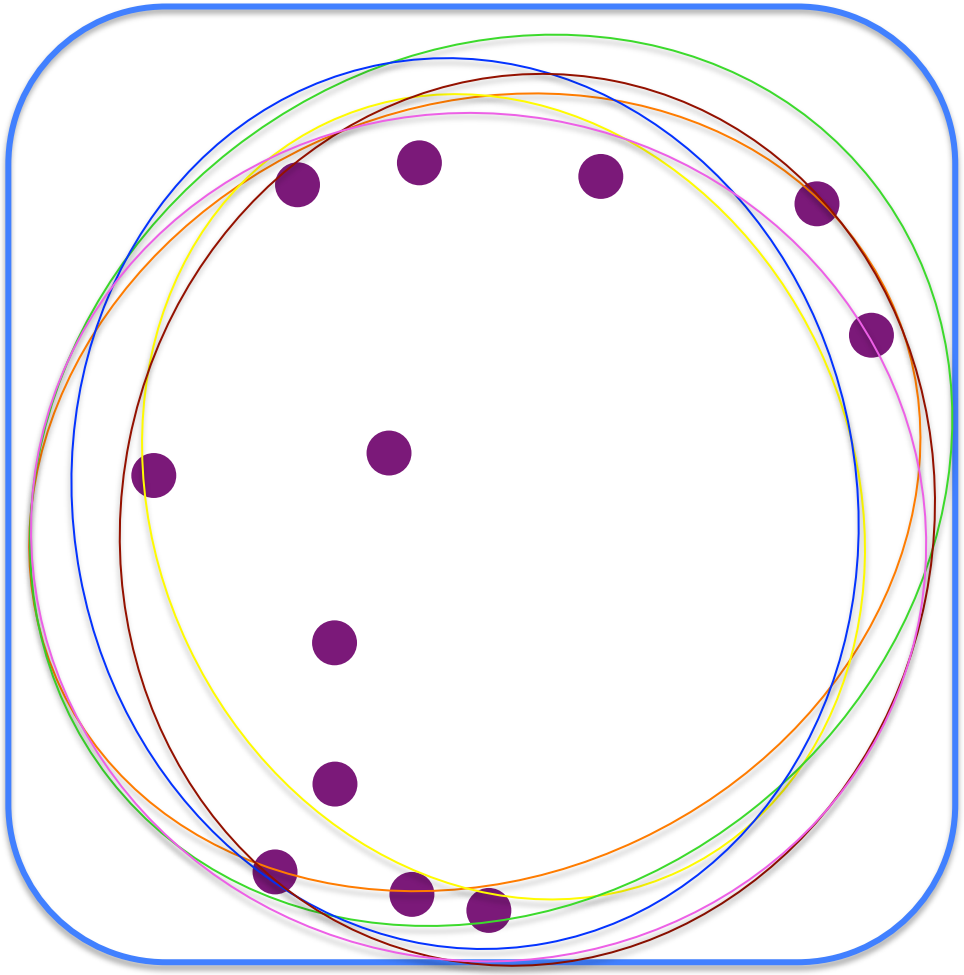
Sequence space

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 …

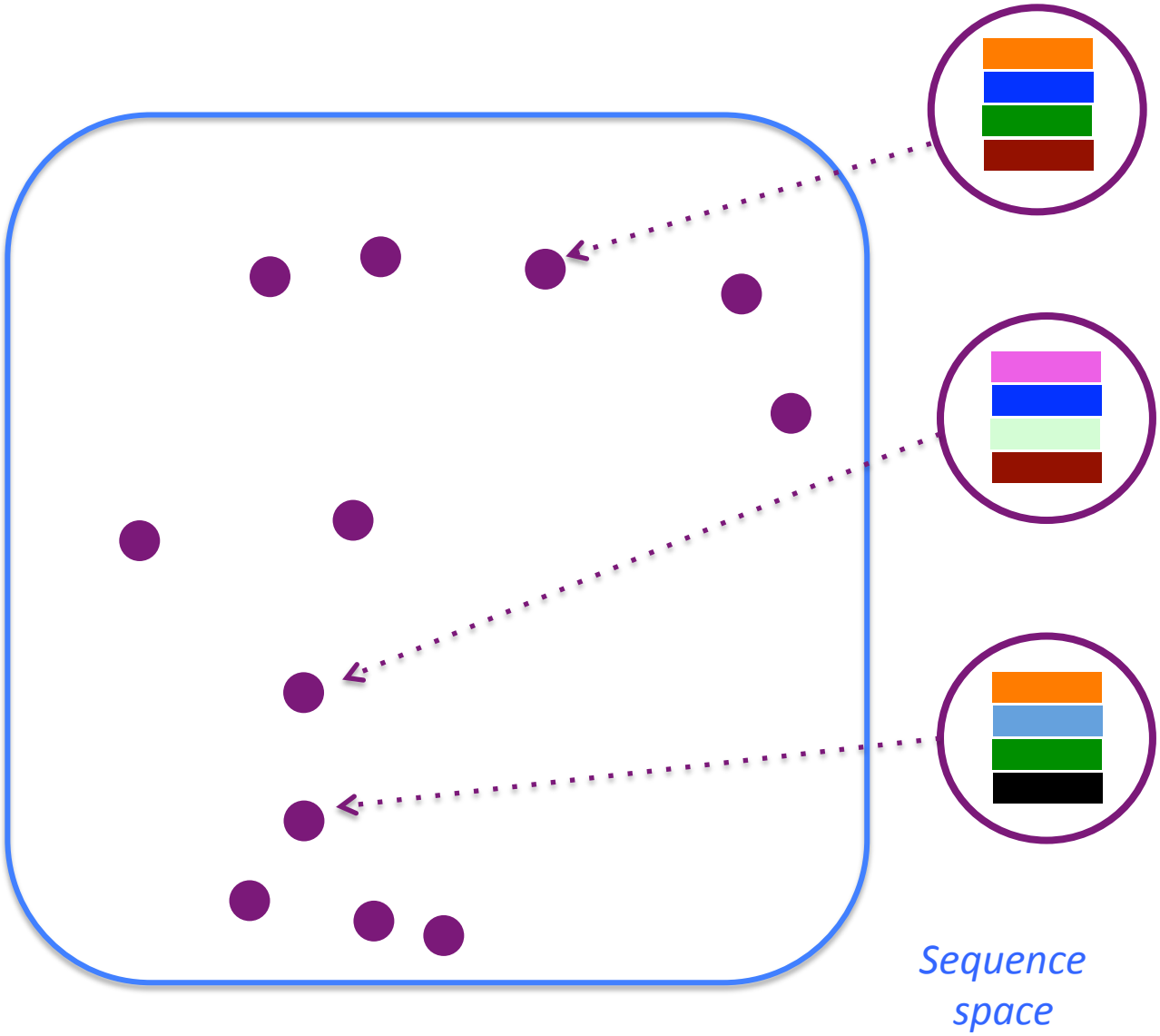


Sequence space

Why are RBM extracting useful/interpretable representations?

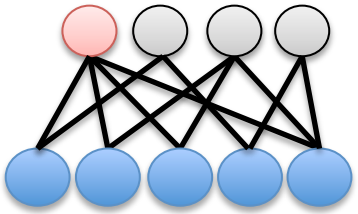
Decomposition into constitutive features:

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



The three representational regimes of RBM

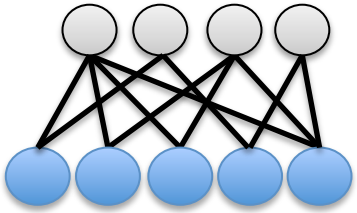
"Ferromagnetic" regime



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

More hidden units

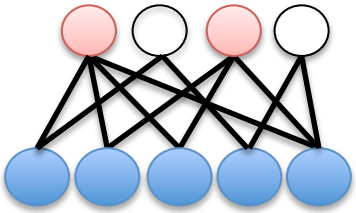
"Glassy" regime



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

Increasing sparsity

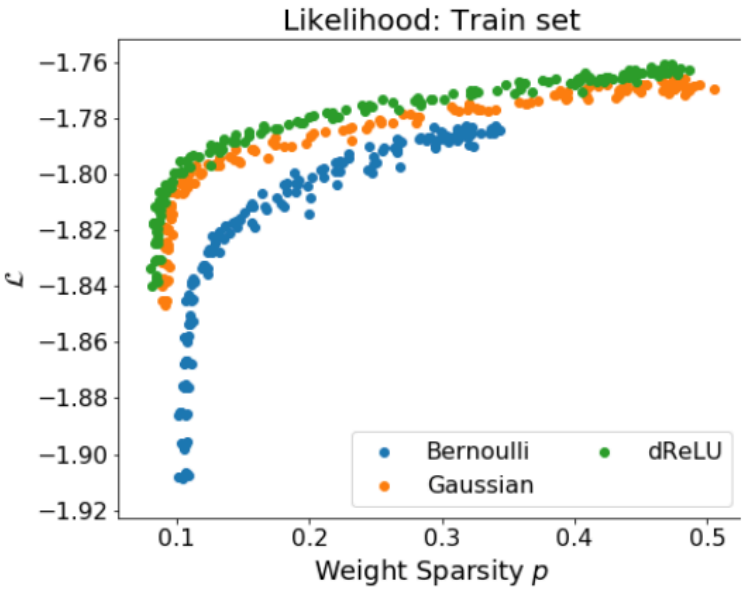
Non quadratic hidden-unit potentials



Compositional regime

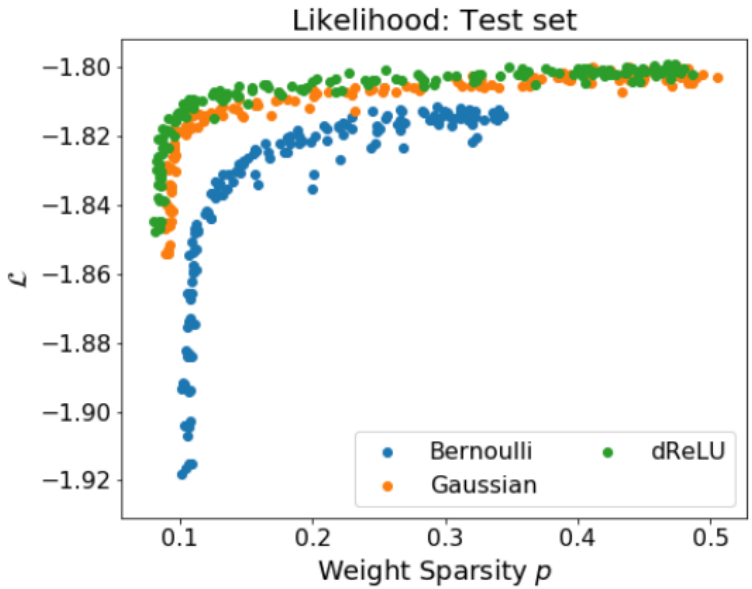
- Multiple hidden units very active
- Corresponding weights define features composing visible configurations
- Extract invariances; interpretable

Driving RBM to the compositional phase



Very sparse

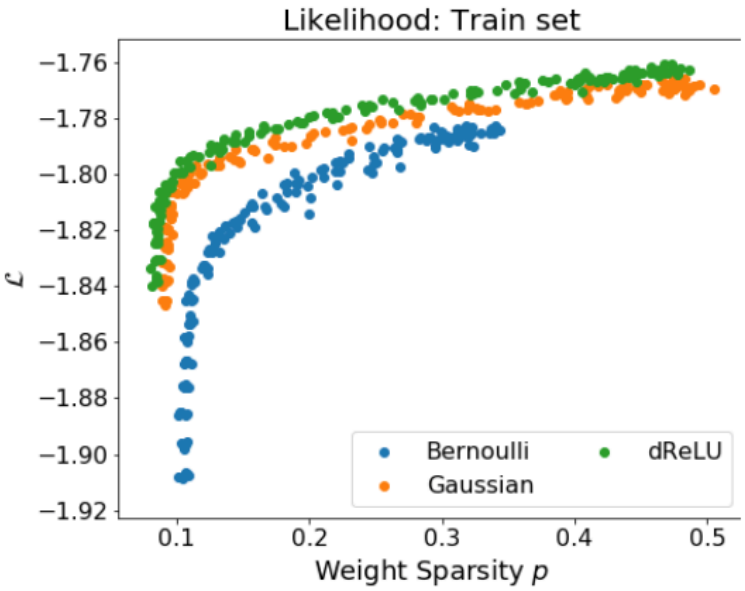
Not sparse

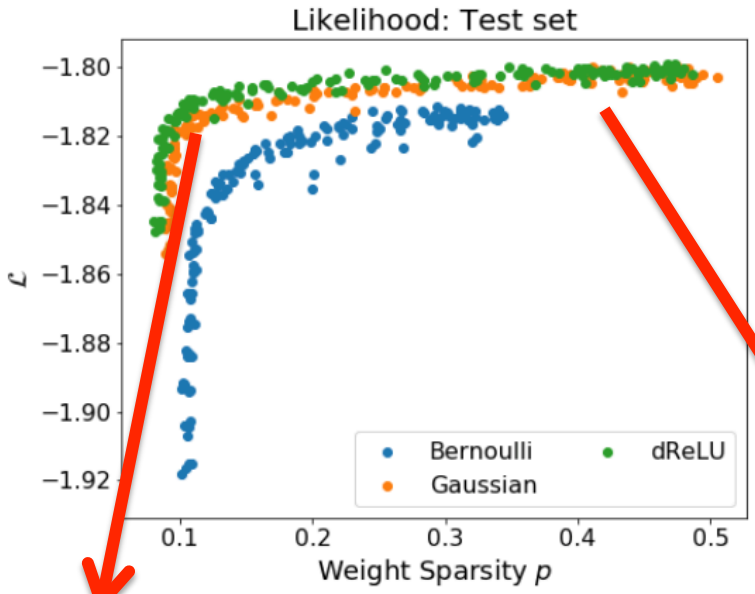


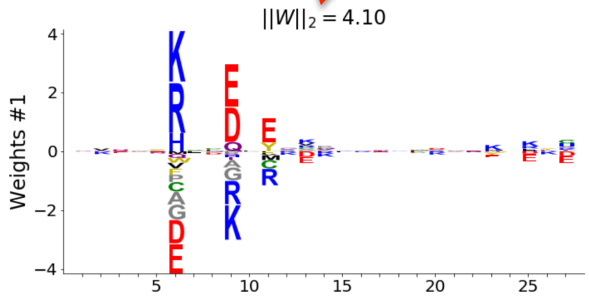
Very sparse

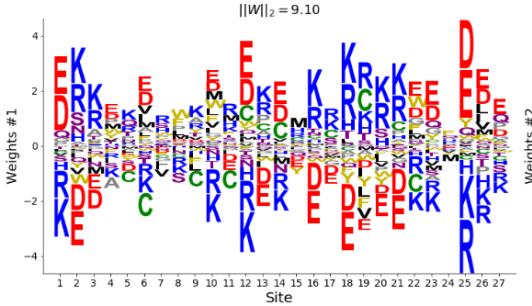
Not sparse

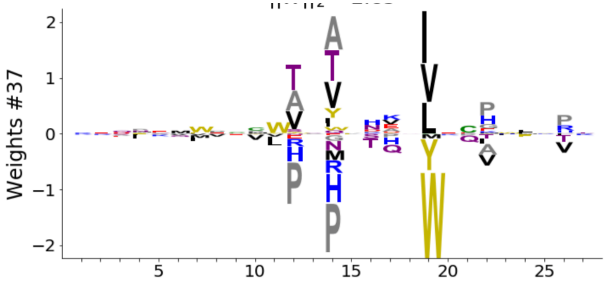
Driving RBM to the compositional phase

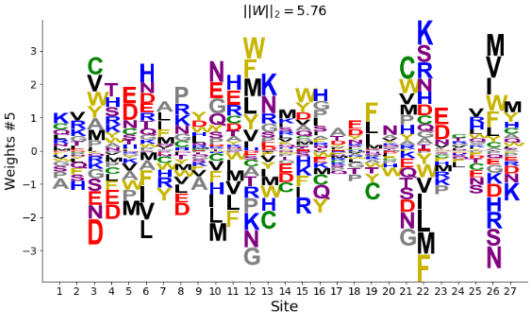








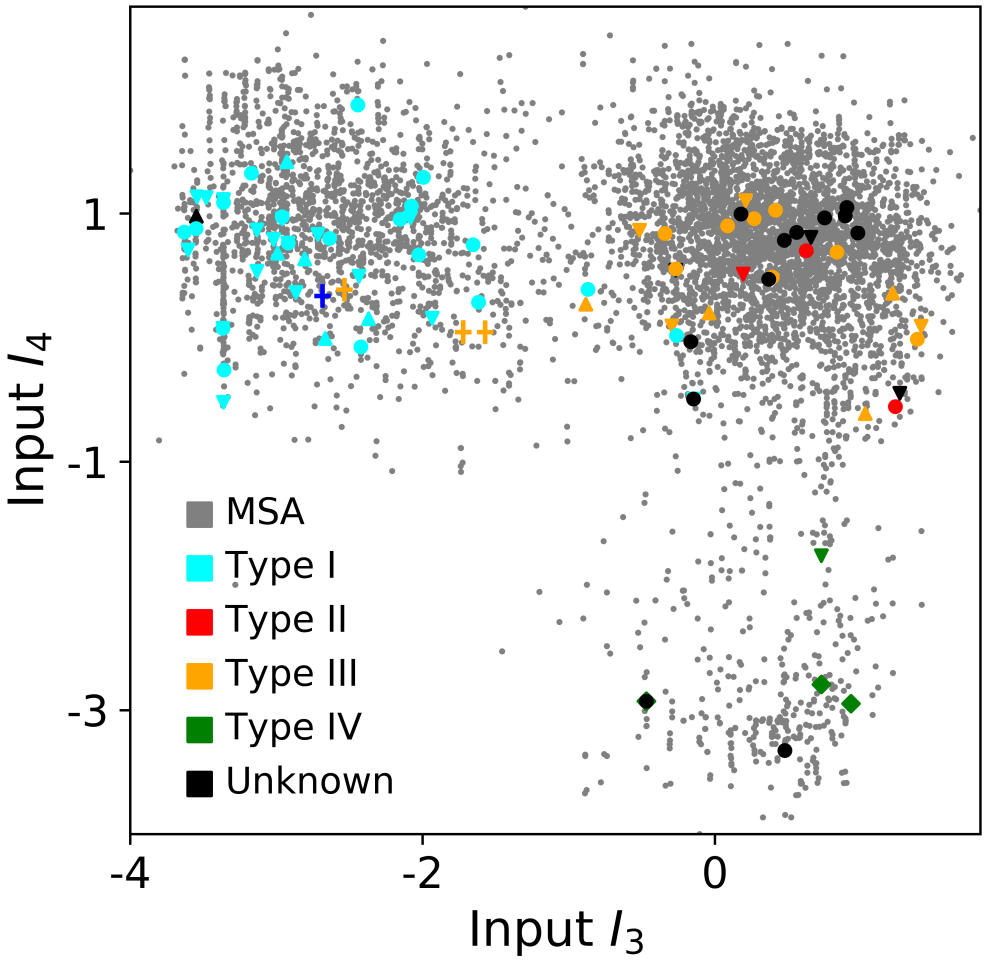




WW Domain: Weights Determine Motif Recognition Specificity

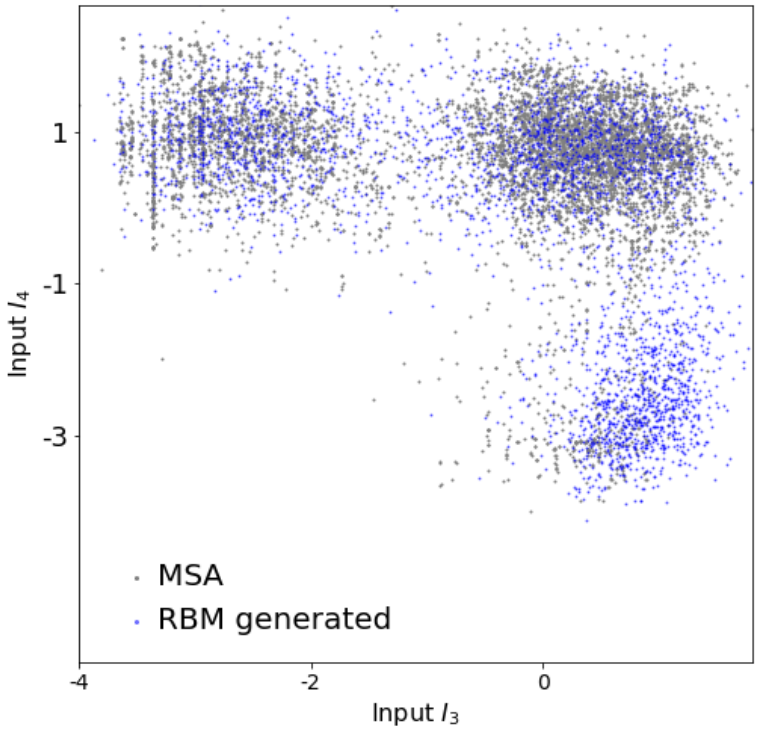
Motif recognized

Type I : PPXY
Type II : PPLP
Type III: PR
Type IV: [p(S/T)P]



Experimental data from:
Russ et al. Nature 2005
Espanel and Sudol J. Biol. Chem. 1999
Otte et al. Protein Science 2003

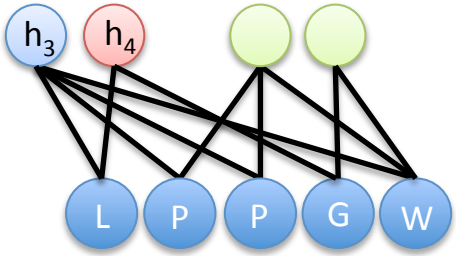
Artificial Sequence Generation with RBM

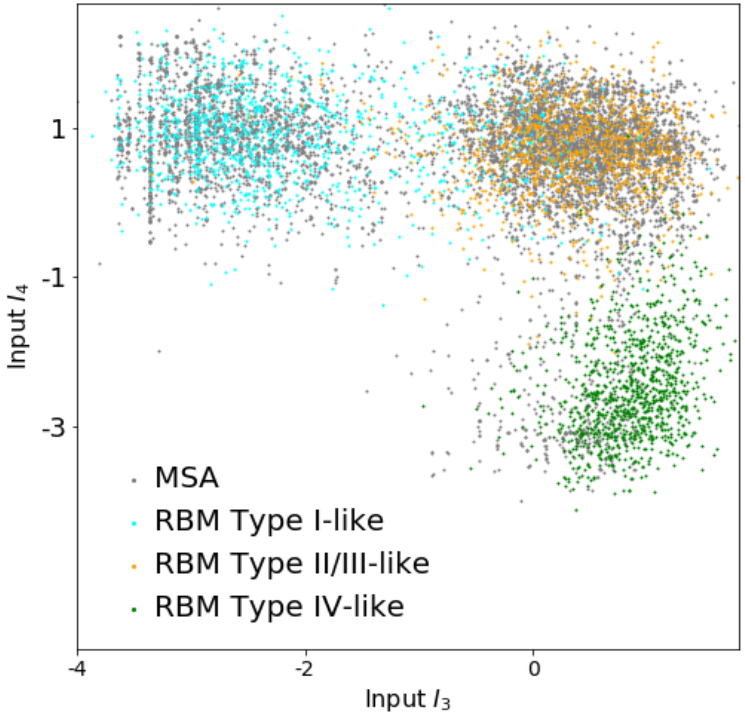


Artificial Sequences

Artificial Sequence Generation with RBM

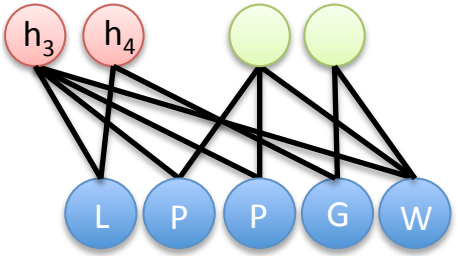
Type I-like
binding pocket +
Short loop



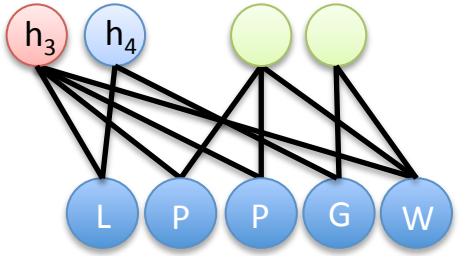


Artificial Sequences

Type II-like
binding pocket +
Short loop

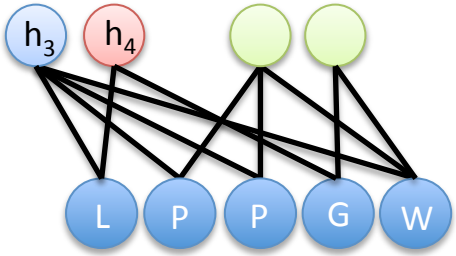


Type II-like
binding pocket +
Short loop

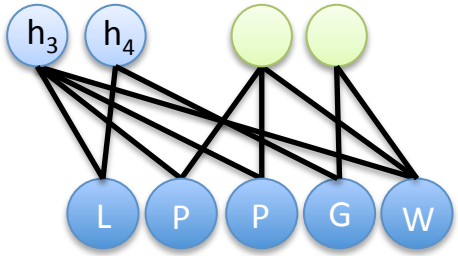


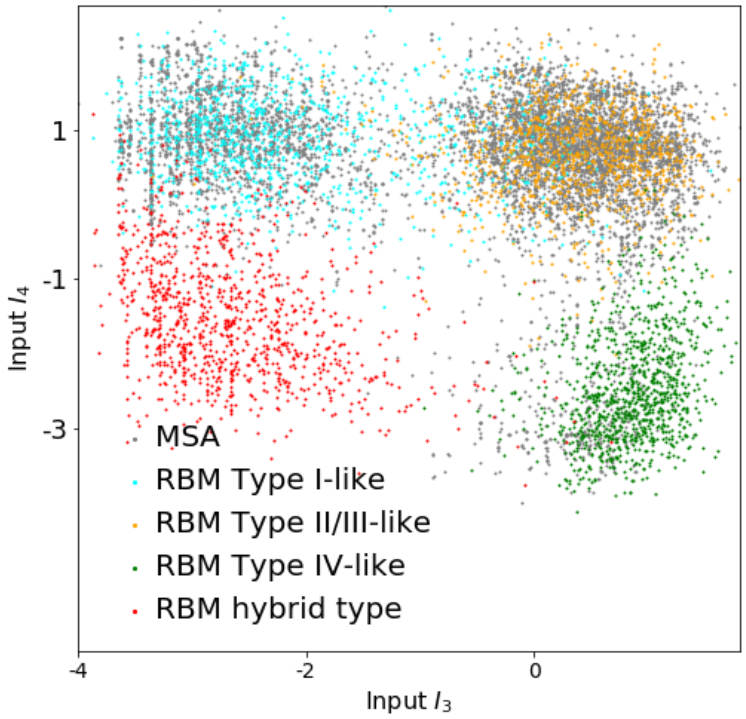
Artificial Sequence Generation with RBM

Type I-like
binding pocket +
Short loop → Type I



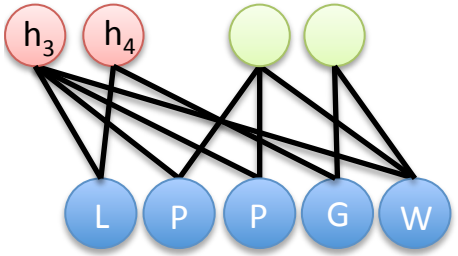
Type I-like
binding pocket +
Short loop → Unknown



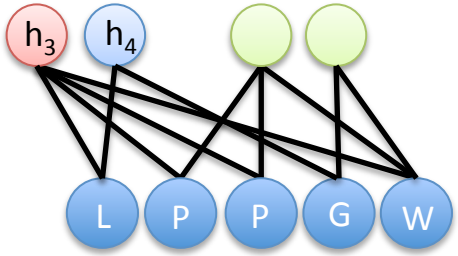


Artificial Sequences

Type II/III/IV-like
binding pocket +
Short loop → Type II/III



Type II/III/IV-like
binding pocket +
Short loop → Type IV



Perspectives

- Data-driven models are getting increasingly important; need for controlled approaches to infer and interpret models.

- If you have questions about lectures or notes, please contact me!
  (remi.monasson at ens.fr)

- Fascinating issues from statistical, computational, conceptual points of views. Lots of things to do in future!!

- Postdoc positions available from January 2020 …