Extracting features from protein sequence data with Restricted Boltzmann Machines

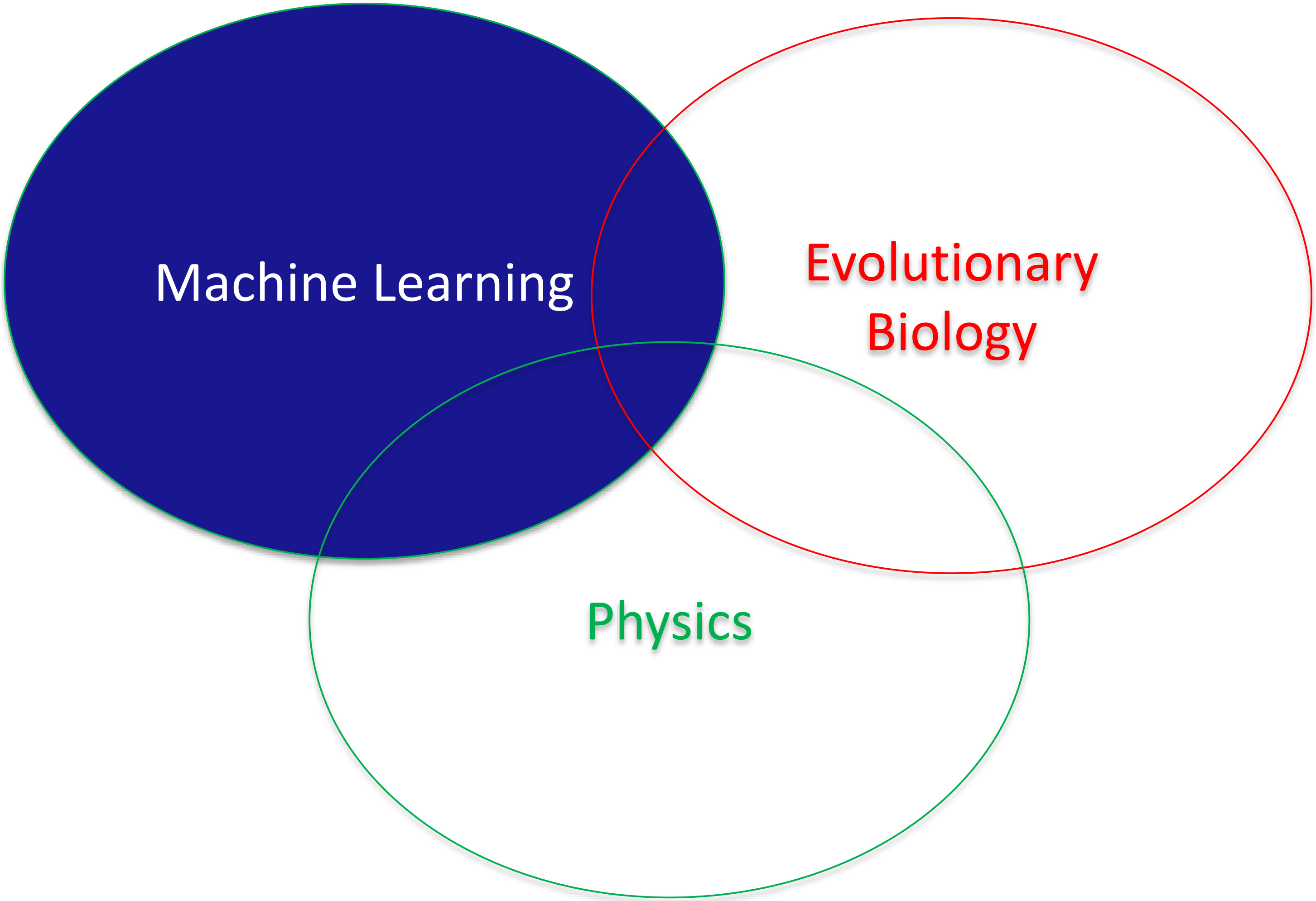
Rémi Monasson, Jérôme Tubiana

Laboratory for Theoretical Physics, Ecole Normale Superieure & CNRS, Paris

Simona Cocco

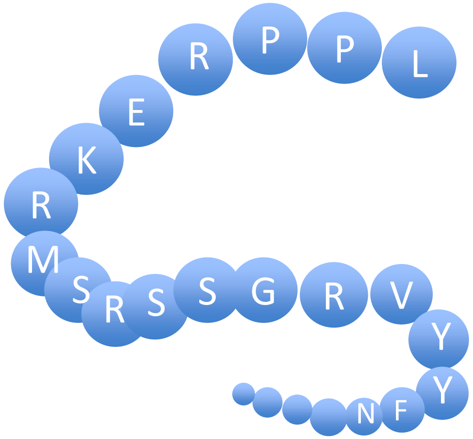
Laboratory for Statistical Physics, Ecole Normale Superieure & CNRS, Paris

The Artificial Intelligence and Physics Conference, Orsay, march 22, 2019


Machine Learning
Evolutionary Biology
Physics

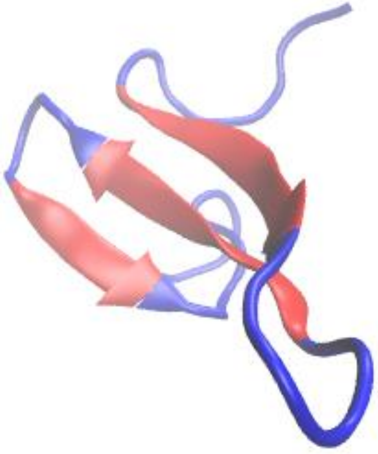
Proteins: from sequence to function

Sequence



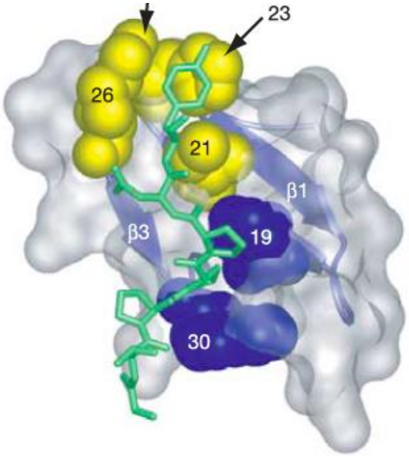
Folding

Structure



Docking

Function

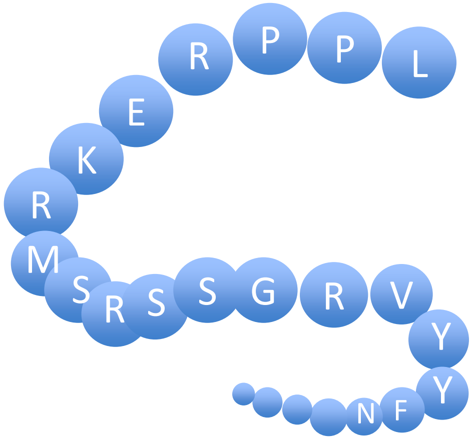


WW domain

[Russ et al. 2005]

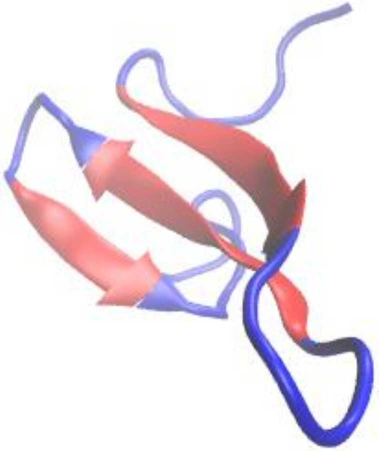
Proteins: from sequence to function

Sequence



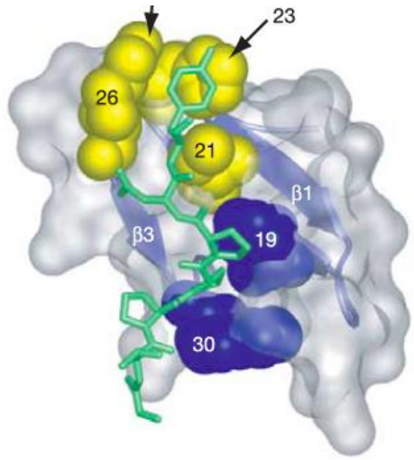
Folding

Structure



Docking

Function

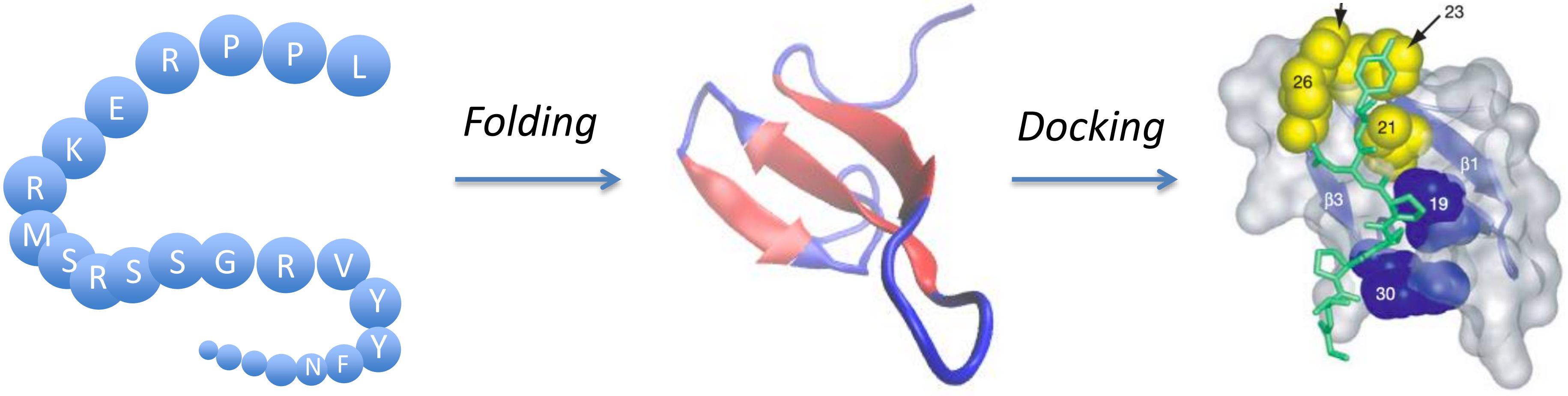


WW domain

[Russ et al. 2005]

VERY HARD FROM
FIRST PRINCIPLE METHODS
(Molecular Dynamics, Ab Initio…)

Constraints on protein sequences



Sequence

Folding

Structure

Docking

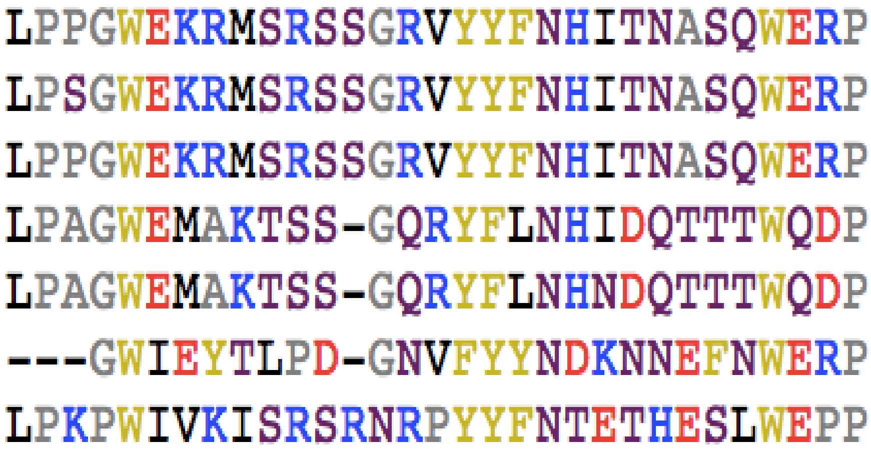
Function

WW domain

[Russ et al. 2005]

- **Stability:** Must fold, and only in the native(s) fold(s)
- **Affinity:** Must bind to target ligand
- **Specificity:** Must preferentially bind a specific ligands…
- **Catalytic Activity:** Must promote reaction within the target ligand
- **Allostery:** Must change conformation upon partner binding…

Proteins: from sequence to function

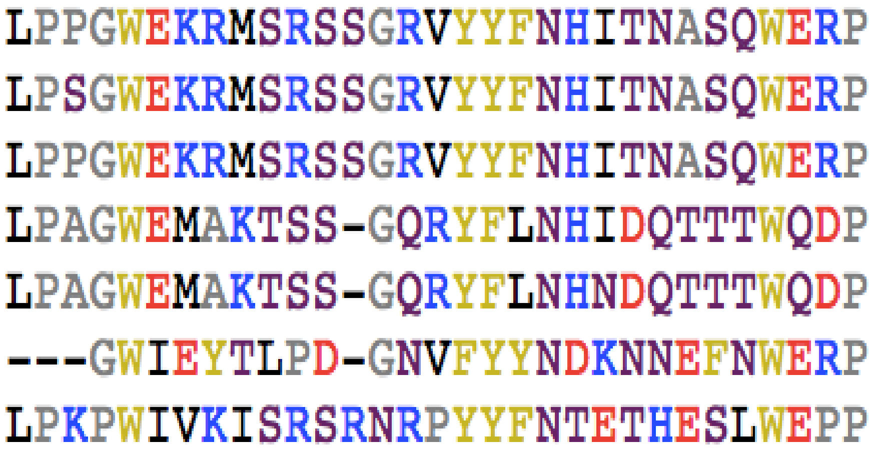


```
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPSGWEKRMSRSSGRVYYFNHITNASQWERP
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRYFLNHNDQTTTWQDP
---GWIEYTLPD-GNVFYYNDKNNEFNWERP
LPKPWIVKISRSRNRPYYFNTETHESLWEPP
```

Multiple Sequence Alignment of Functional WW-domain sequences from diverse organism and genes

(Source: PFAM)

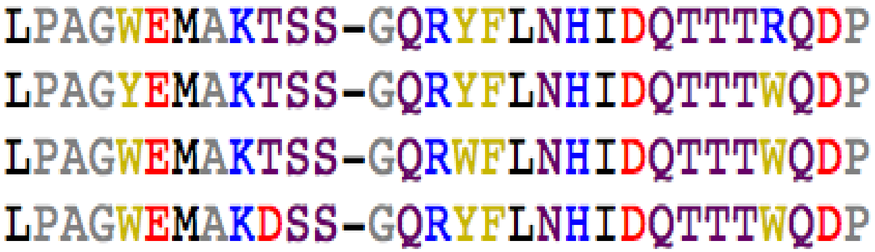
Proteins: from sequence to function



```
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPSGWEKRMSRSSGRVYYFNHITNASQWERP
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRYFLNHNDQTTTWQDP
---GWIEYTLPD-GNVFYYNDKNNEFNWERP
LPKPWIVKISRSRNRPYYFNTETHESLWEPP
```

Functional WW-domain sequences from diverse organism and genes

(Source: PFAM)

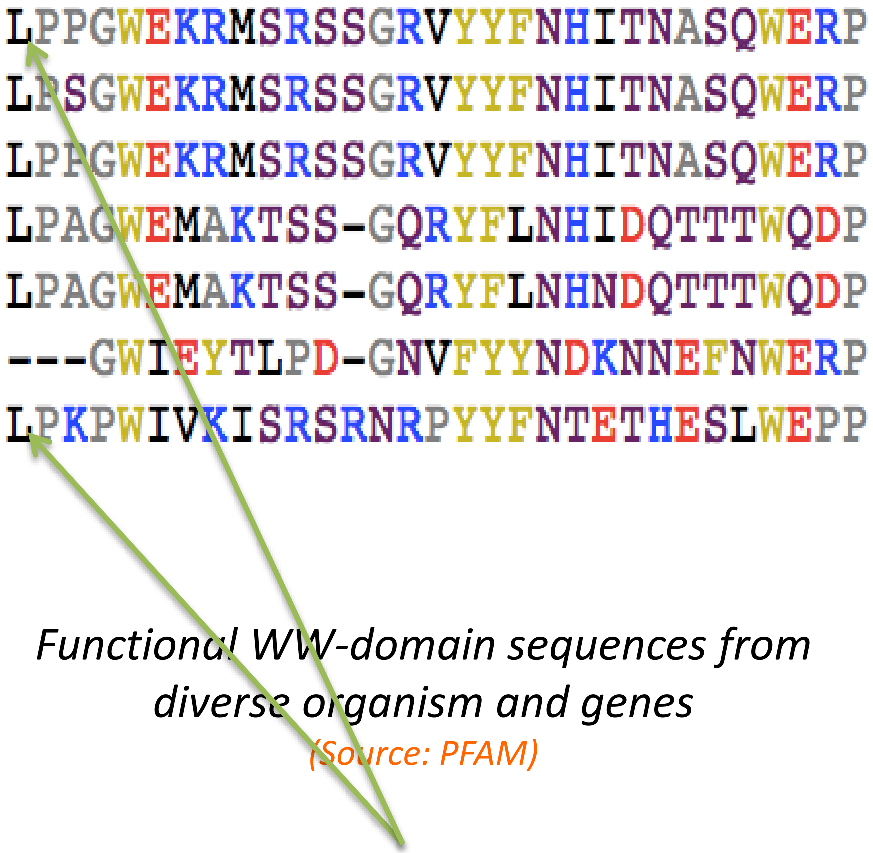


```
LPAGWEMAKTSS-GQRYFLNHIDQTTTRQDP
LPAGYEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRWFLNHIDQTTTWQDP
LPAGWEMAKDSS-GQRYFLNHIDQTTTWQDP
```

Examples of non-functional WW sequences obtained by mutagenesis

(Fowler et al. Nature Methods 2011)

Proteins: from sequence to function



```
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPSGWEKRMSRSSGRVYYFNHITNASQWERP
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRYFLNHNDQTTTWQDP
---GWIEYTLPD-GNVFYYNDKNNEFNWERP
LPKPWIVKISRSRNRPYYFNTETHESLWEPP
```

Functional WW-domain sequences from diverse organism and genes

(Source: PFAM)

<50% Sequence identity.

Same activity *in vitro*

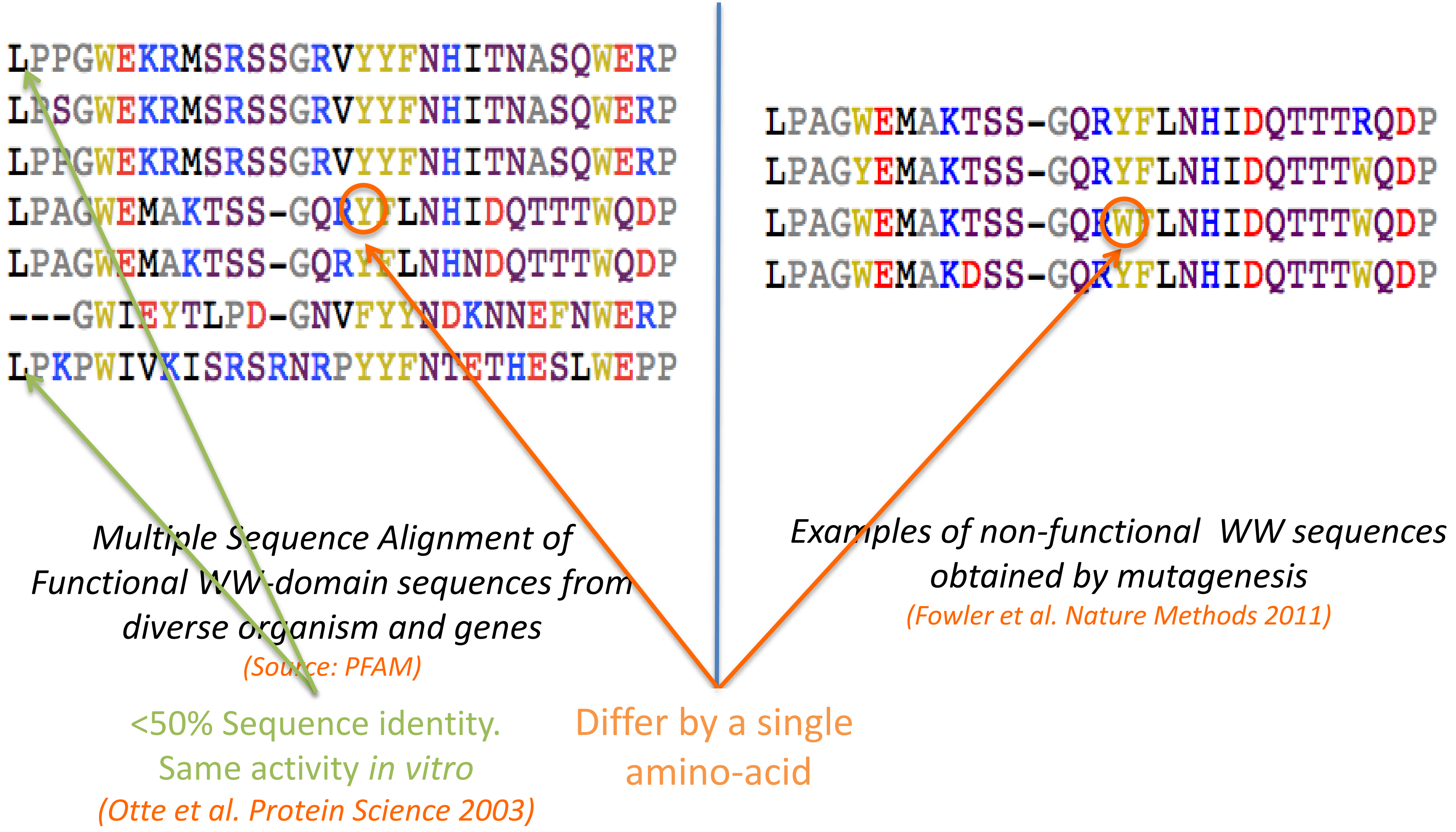
(Otte et al. Protein Science 2003)

```
LPAGWEMAKTSS-GQRYFLNHIDQTTTRQDP
LPAGYEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRWFLNHIDQTTTWQDP
LPAGWEMAKDSS-GQRYFLNHIDQTTTWQDP
```

Examples of non-functional WW sequences obtained by mutagenesis

(Fowler et al. Nature Methods 2011)

Proteins: from sequence to function



```
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPSGWEKRMSRSSGRVYYFNHITNASQWERP
LPPGWEKRMSRSSGRVYYFNHITNASQWERP
LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRYFLNHNDQTTTWQDP
---GWIEYTLPD-GNVFYYNDKNNEFNWERP
LPKPWIVKISRSRNRPYYFNTETHESLWEPP
```

Multiple Sequence Alignment of Functional WW-domain sequences from diverse organism and genes

(Source: PFAM)

```
LPAGWEMAKTSS-GQRYFLNHIDQTTTRQDP
LPAGYEMAKTSS-GQRYFLNHIDQTTTWQDP
LPAGWEMAKTSS-GQRWFLNHIDQTTTWQDP
LPAGWEMAKDSS-GQRYFLNHIDQTTTWQDP
```

Examples of non-functional WW sequences obtained by mutagenesis

(Fowler et al. Nature Methods 2011)

<50% Sequence identity.
Same activity *in vitro*
(Otte et al. Protein Science 2003)

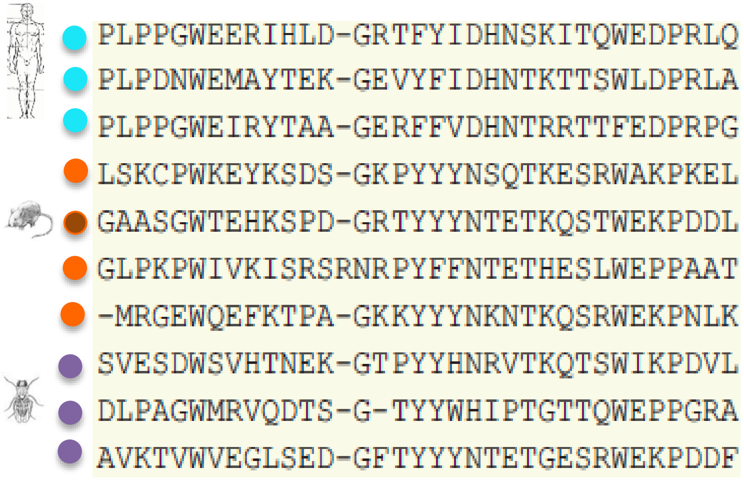
Differ by a single amino-acid

Proteins: from sequence to function

- What makes a protein sequence functional ?
- Can we find biologically relevant representations of these sequences ?
- Can we design functional artificial sequences ?

Mutational Landscape of Proteins

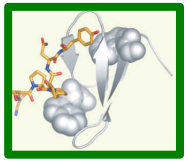
Data: Multi Sequence Alignment (MSA)

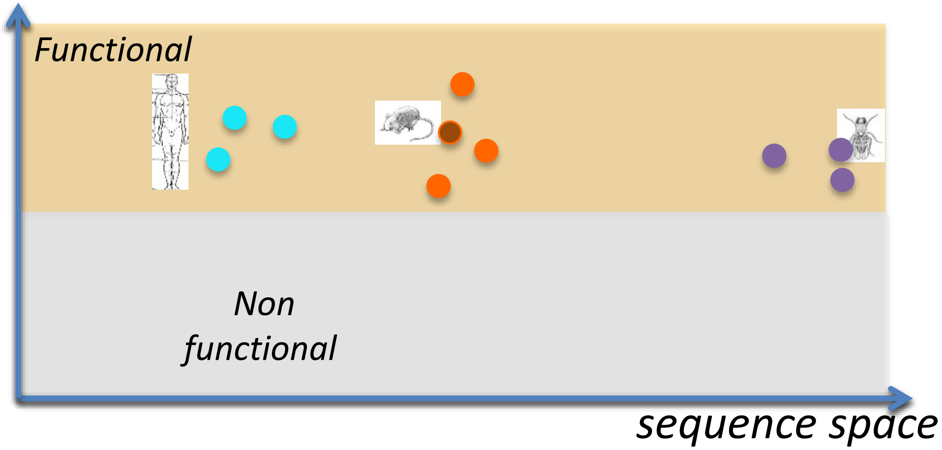


```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```

(Source: PFAM)

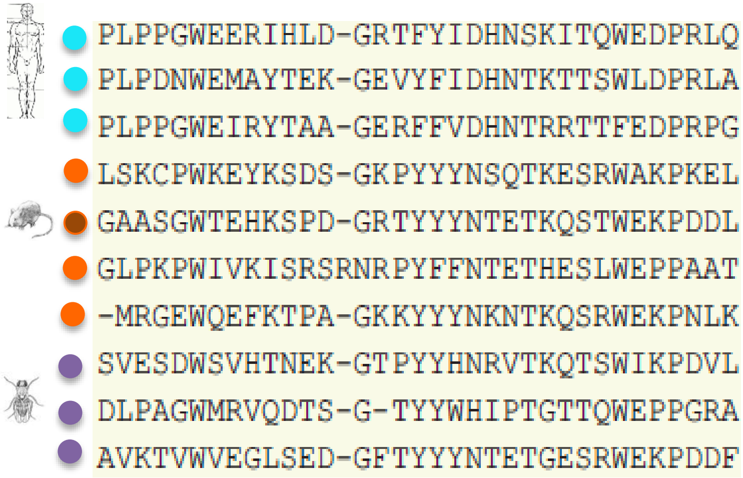
Mutational Landscape





Mutational Landscape of Proteins

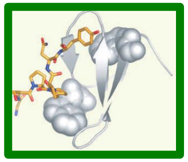
Data: Multi Sequence Alignment (MSA)

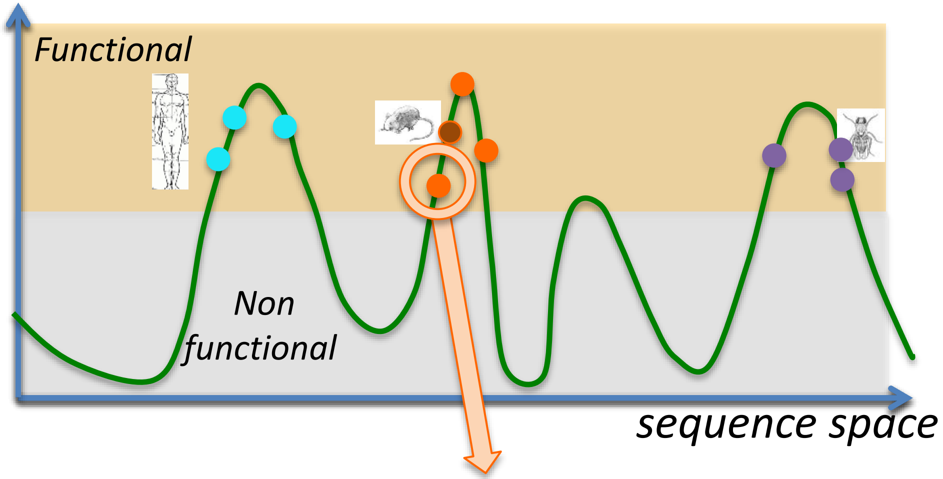


```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```

*(Source: PFAM, typically 10^3-10^5*sequencces*)*

Model: Probability of a sequence





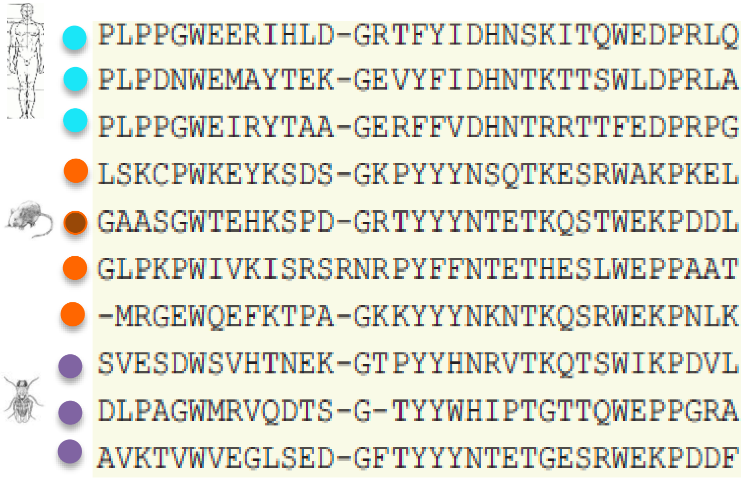
Prediction for change in functionality due to single, double, … mutations

Infer from the data the Probability of a sequence $P(v_1....v_N)$ to be a good sequence for that protein.

v_i =A,C,D….W,-, are the 20 amino acid of the protein +gap symbol
Potts (categorigal) variables

Mutational Landscape of Proteins

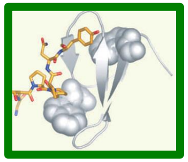
Data: Multi Sequence Alignment (MSA)

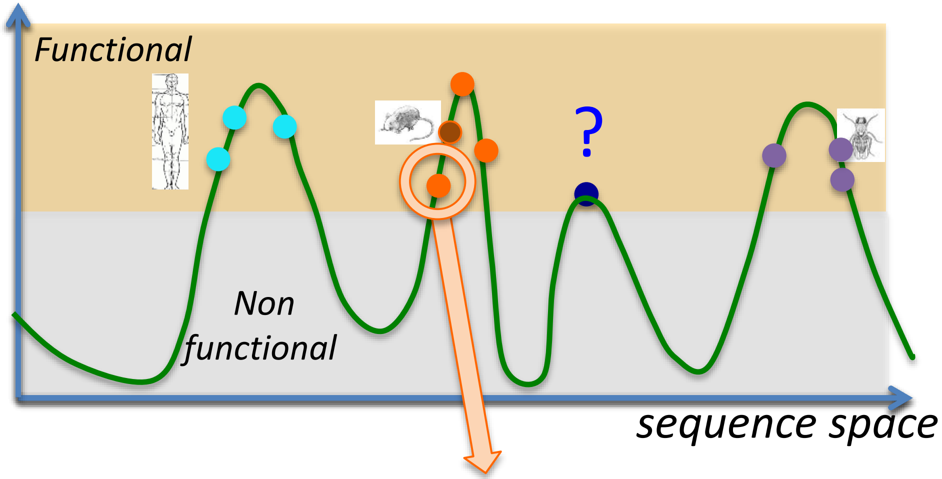


PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF

(Source: PFAM, typically 10^3-10^5sequencces)

Model: Probability of a sequence



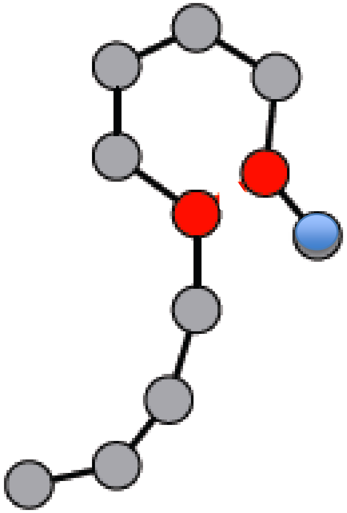


Prediction for change in functionality due to single, double, … mutations **& design new sequences**

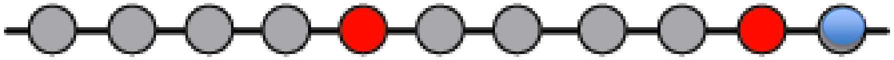
Infer from the data the Probability of a sequence $P(v_1....v_N)$ to be a good sequence for that protein.

v_i =A,C,D….W,-, are the 20 amino acid of the protein +gap symbol
Potts (categorigal) variables

Model inference from data



Structural, functional constraints



R	I	D	G	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	E	E	V	E	N	A	K	Q
R	L	N	G	R	A	D	D	L	D	H

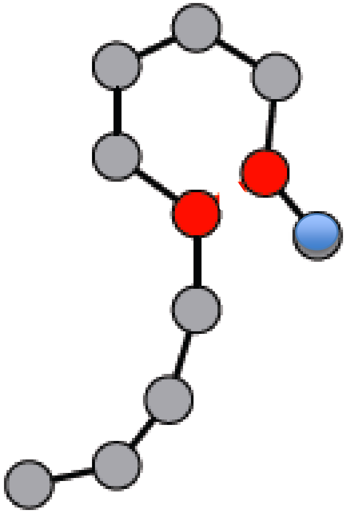
Correlation $f_{ij}(v_i,v_j)$

Covariation

Frequency $f_i(v_i)$:

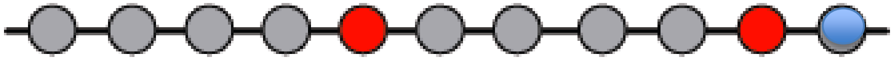
Conservation

Model inference from data



Structural, functional constraints

Inverse Statistical Modeling



R	I	D	G	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	E	E	V	E	N	A	K	Q
R	L	N	G	R	A	D	D	L	D	H

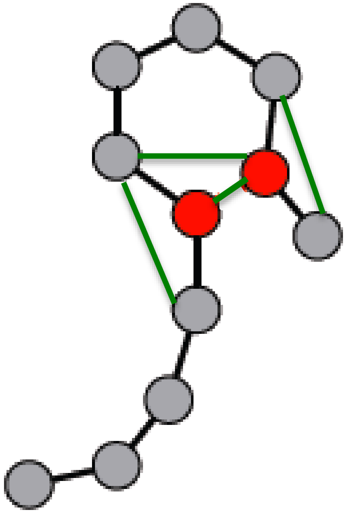
Correlation f_{ij} (v_i,v_j)

Covariation

Frequency f_i (v_i):

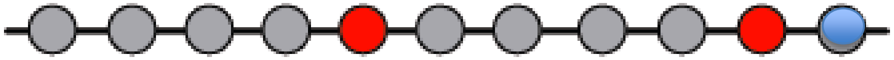
Conservation

Network inference from data



Structural, functional constraints →

← Inverse Statistical Modeling



R	I	D	G	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	E	E	V	E	N	A	K	Q
R	L	N	G	R	A	D	D	L	D	H

Correlation $f_{ij}(v_i,v_j)$
Covariation

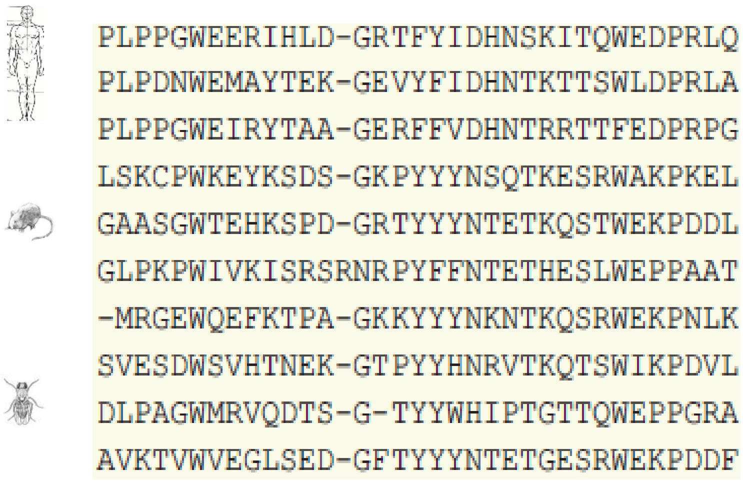
Frequency $f_i(v_i)$:
Conservation

Least constrained, maximal entropy model (Jaynes 1957) reproducing frequencies $f_i(v_i)$ and correlations $f_{ij}(v_i,v_j)$ of empirical distribution

$$P(v_1,\dots,v_N) = \frac{e^{\sum_i h_i(v_i) + \sum_{i<j} J_{ij}(v_i,v_j)}}{Z[\{\mathbf{J},\mathbf{h}\}]}$$

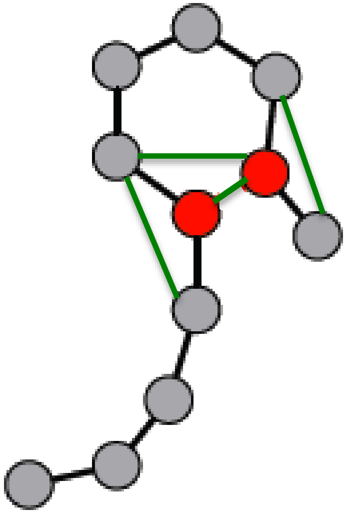
[Morcos … Weigt, PNAS 2011 , Ekeberg, Aurell (2015), Hopf, Colwell et al Cell (2012),Baker(2014),S.C.…Weigt(2017)]

Network inference from data



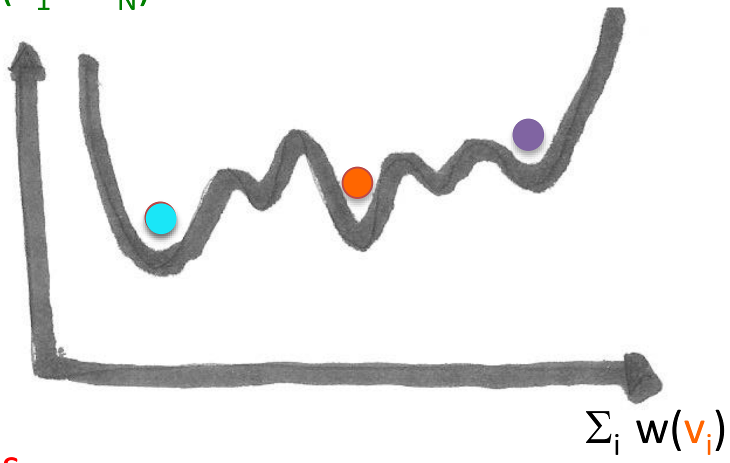
```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```

$J_{ij}\,(v_i,v_j)$



Generate New Sequences
by Monte Carlo simulations

Energy= $-\log P(v_1 \ldots v_N)$

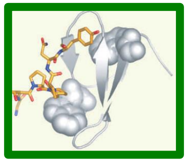


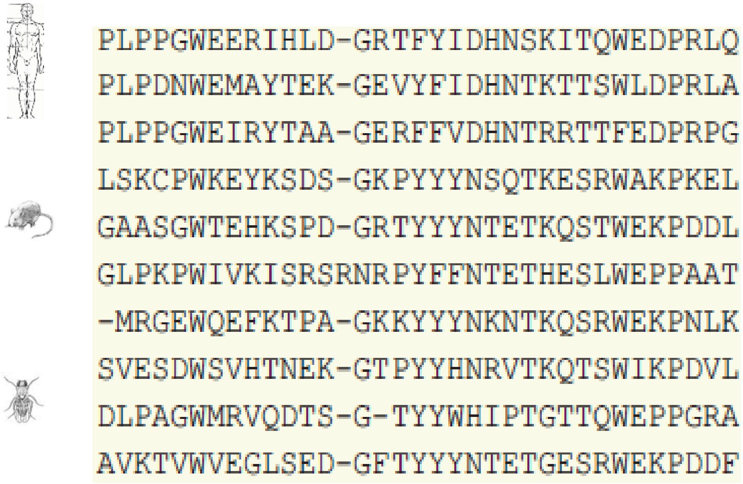
Direct Coupling Analysis:

[Morcos … Weigt, PNAS 2011 , Ekeberg, Aurell (2015)
Hopf, Colwell et al Cell (2012), Baker(2014), S.C. …Weigt(2017)]

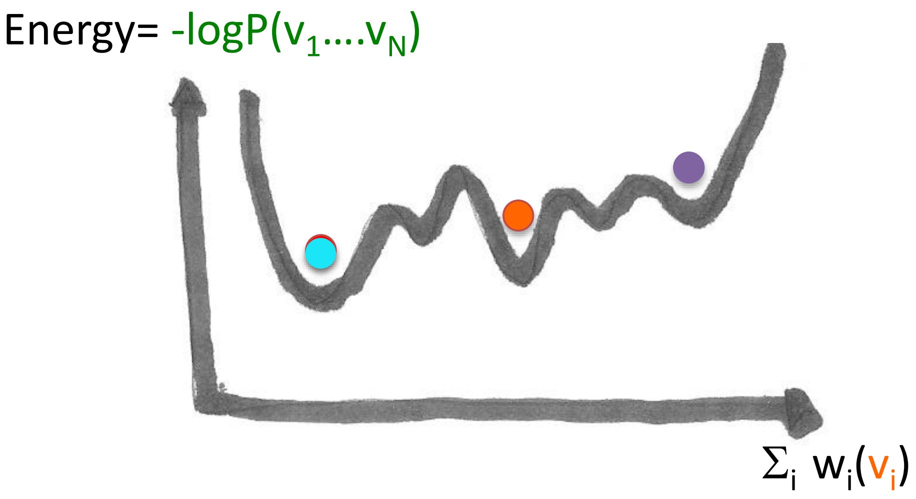
✓Give structural informations
✓Model is generative
✓Predicts cost of mutations and design new sequences
Does not gives direct information on the ‘good’ sequences

Features extraction from data





```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```



PCA, Sparse PCA, and Sector Analysis :

From correlation matrix extract principal components: features $w(v_i)$

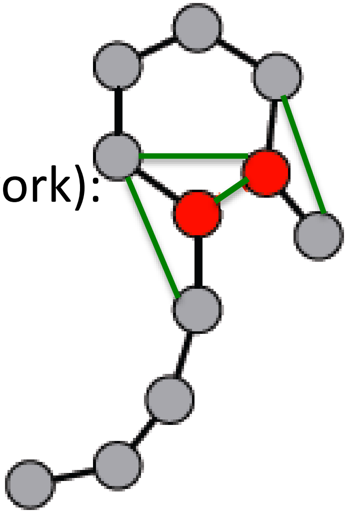
Project sequences on them to characterize the wells of the energy landscape

[A Raussel.. A Valencia 2010, N Halabi,…R.Ranganathan 2009]

But are not generative ..

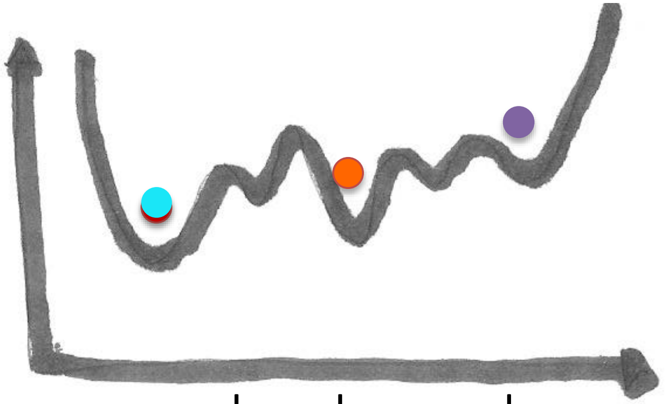
The Hopfield Model

One can built up a coupling matrix storing M features or ’patterns’ as energy minima of the model (associative memories of the network):



$$J_{ij}\,(v_i,v_j) = \sum_{\mu=1}^{M} w_i^{\mu}\,(v_i)\; w_j^{\mu}\,(v_j)$$

$$E(V) = -\frac{1}{2}\sum_{i<j} J_{ij}\,(v_{i,}v_j) \Longrightarrow E(v) = -\frac{1}{2}\sum_{\mu=1}^{M}\Big(\sum_i w_i^{\mu}\,(v_i)\Big)^2$$

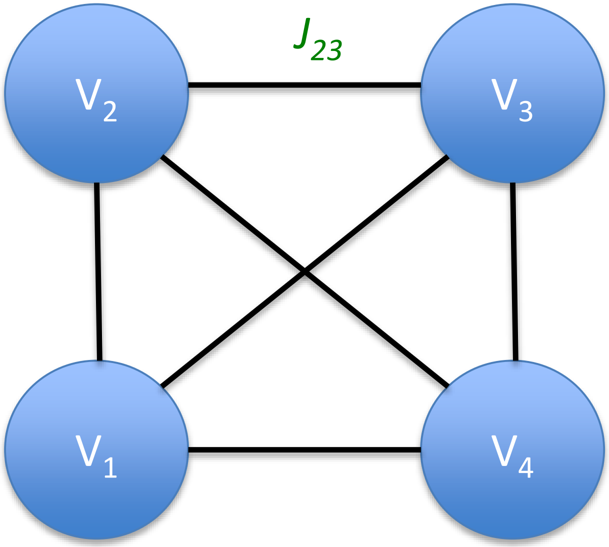


[Hopfield, PNAS 1982]
[SC Monasson Weight Plos Comp Bio 2016]
[Barra, Bernacchia,Santucci, Contucci,Neural Network 2012]

Large probability sequences have large scalar products with the feature -> ‘look like’ it

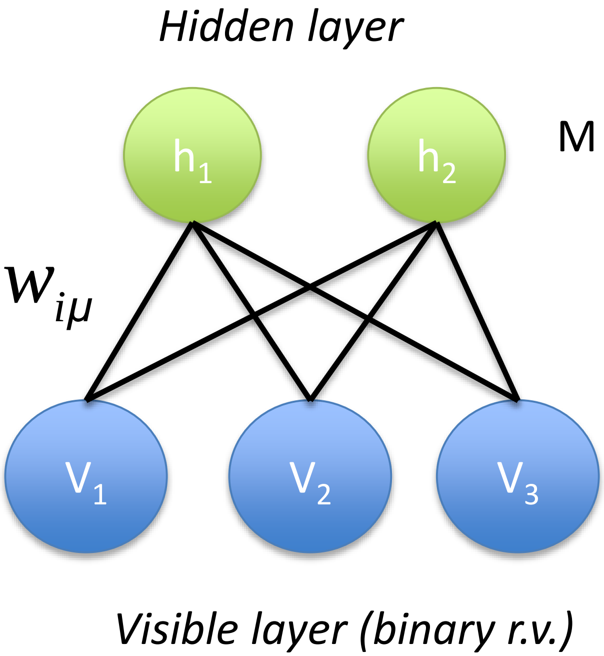
Different Network Architecture

Boltzmann Machine



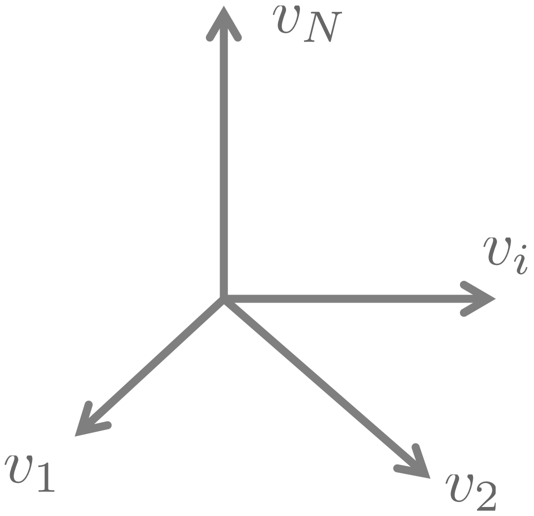
Ising/Potts model (BM) explains correlations by couplings J_{ij} between nodes (variables)

Restricted Boltzmann Machine



RBM explains data through their common features
Combinations of features can, in turn, generate new data

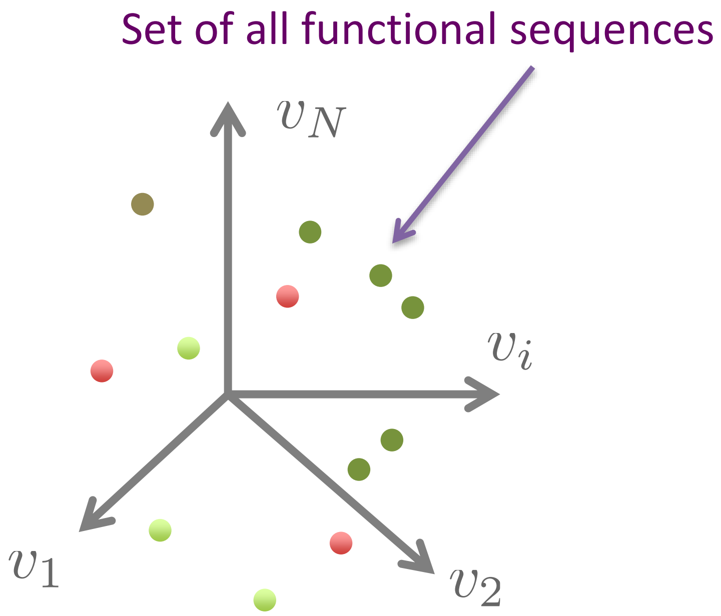
Learning distributions over data



Data space

$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

Learning distributions over data

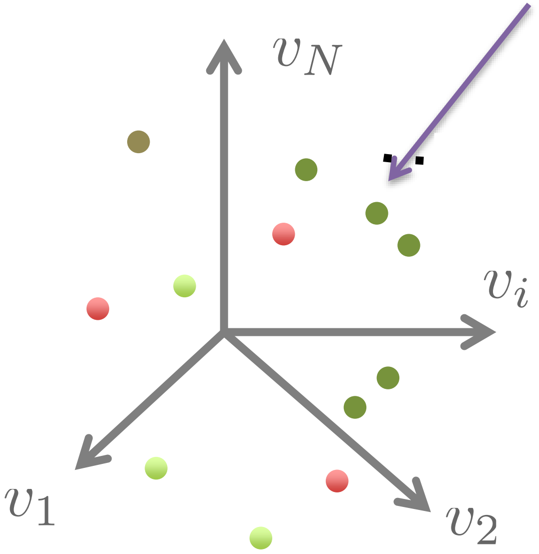


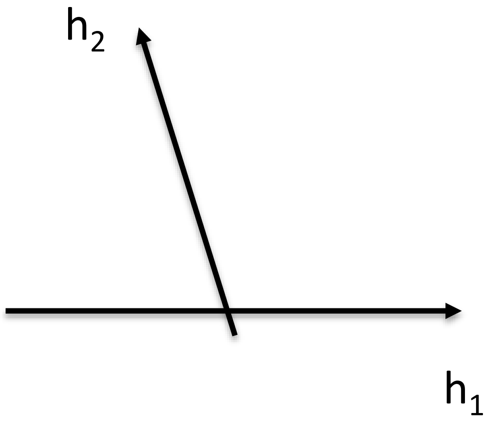
Data space

$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

Learning Representations of data

Set of all functional sequences



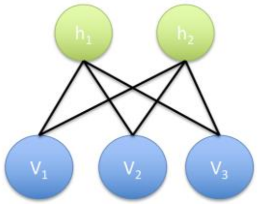


Data space

Representation space

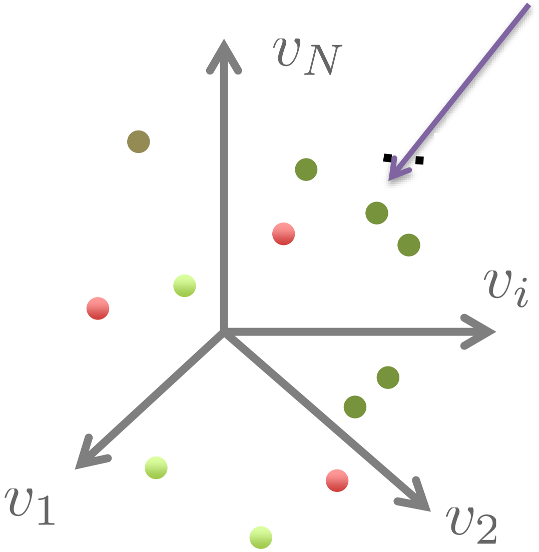
$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

$\{h\}^M$ (features)

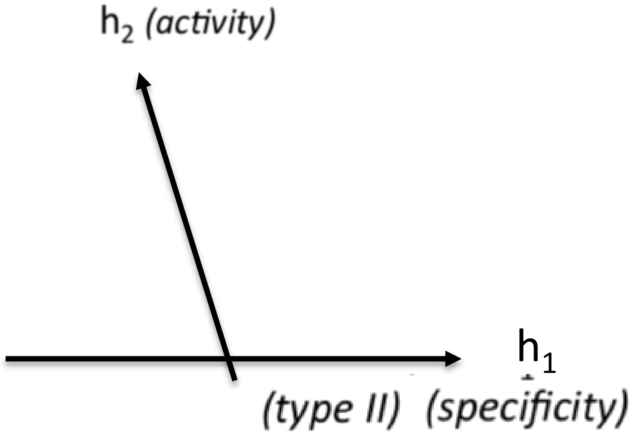


Learning Representations of data

Set of all functional sequences



[Protein bio-chemical properties]

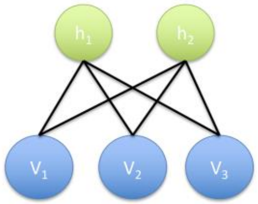


Data space -> Genotype

Representation space ->Phenotype

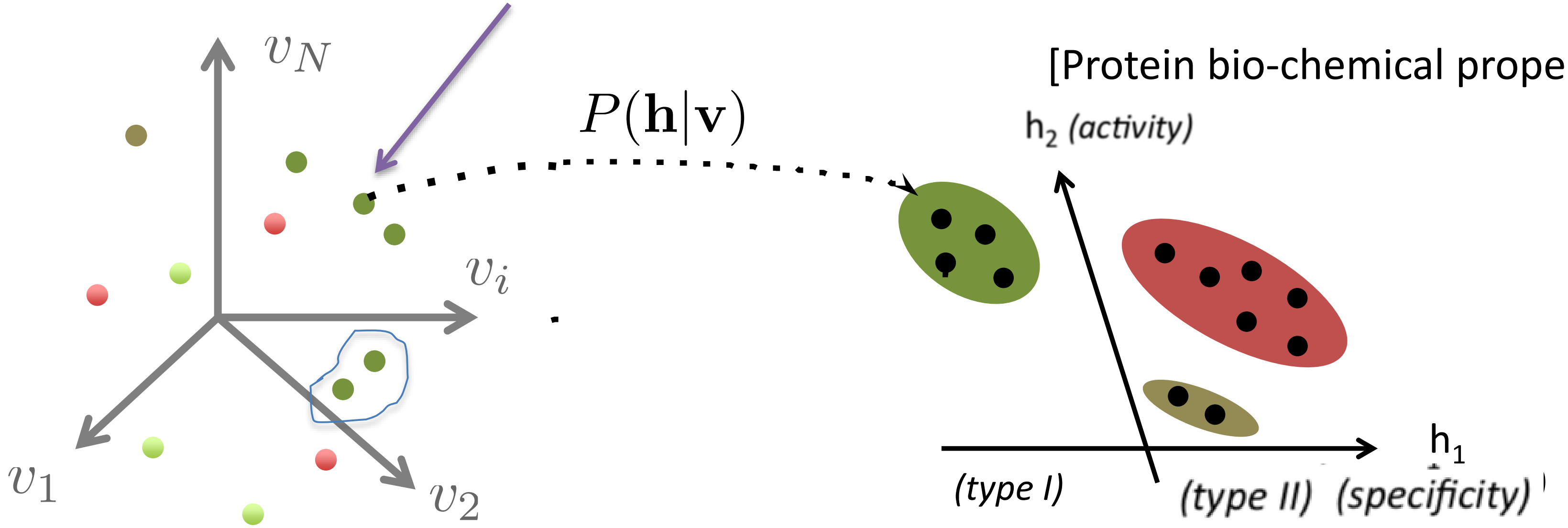
$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

$\{h\}^M$ (features)



Learning Representations of data

Set of all functional sequences

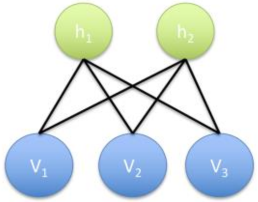


Data space

Representation space

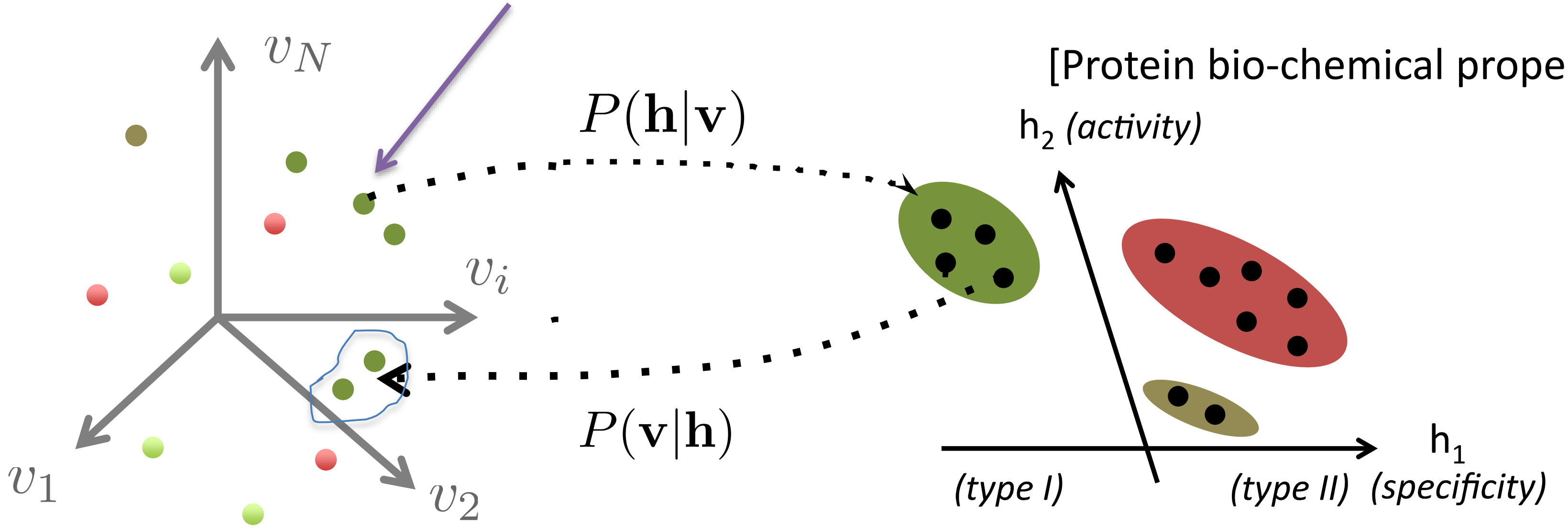
$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

$\{h\}^M$ (features)



Learning Representations of data

Set of all functional sequences



$P(\mathbf{h}|\mathbf{v})$

$P(\mathbf{v}|\mathbf{h})$

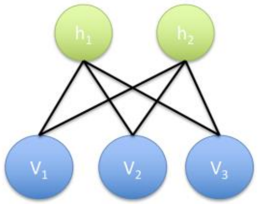
[Protein bio-chemical properties]

Data space

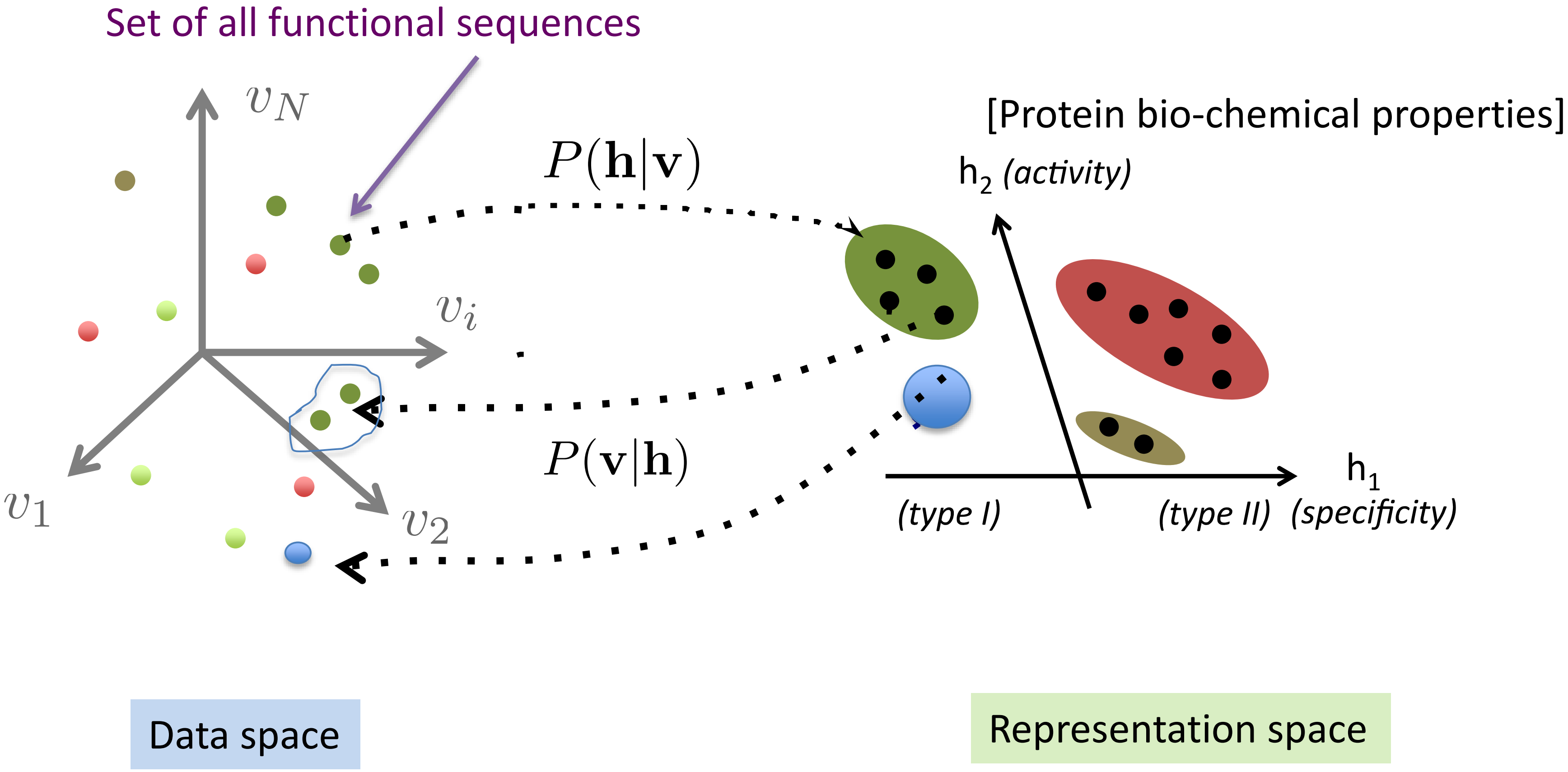
Representation space

$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

$\{h\}^M$ (features)



Learning Representations of data


Set of all functional sequences
v_N
v_i
v_1
v_2
$P(\mathbf{h}|\mathbf{v})$
$P(\mathbf{v}|\mathbf{h})$
[Protein bio-chemical properties]
h_2 *(activity)*
h_1
(type I)
(type II)
(specificity)
Data space
Representation space

$\{A, C, D, E, .., Y, -\}^N$ (protein sequences)

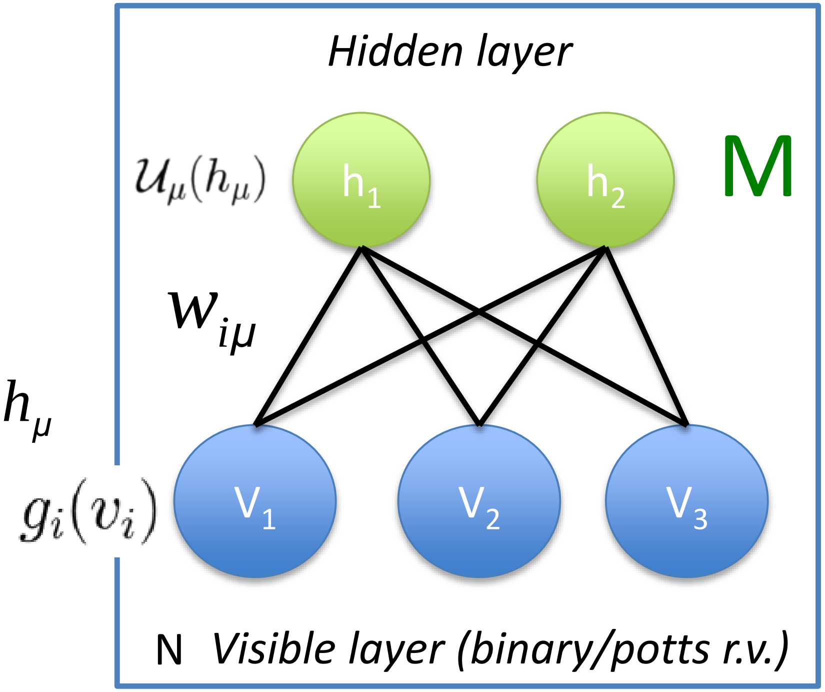
Restricted Boltzmann Machines

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

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

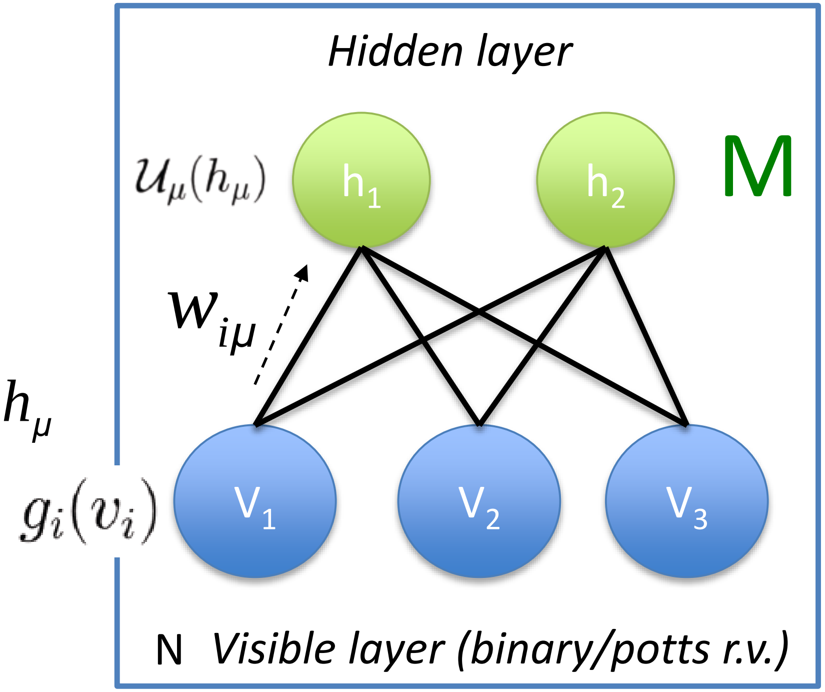
$$E(v,h) = -\sum_{i} g_i(v_i) + \sum_{\mu} U_\mu(h_\mu) - \sum_{i,\mu} w_{i\mu}(v_i)h_\mu$$

.



Restricted Boltzmann Machines

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



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

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

.Given an input configuration, the hidden unit μ Receives an input:

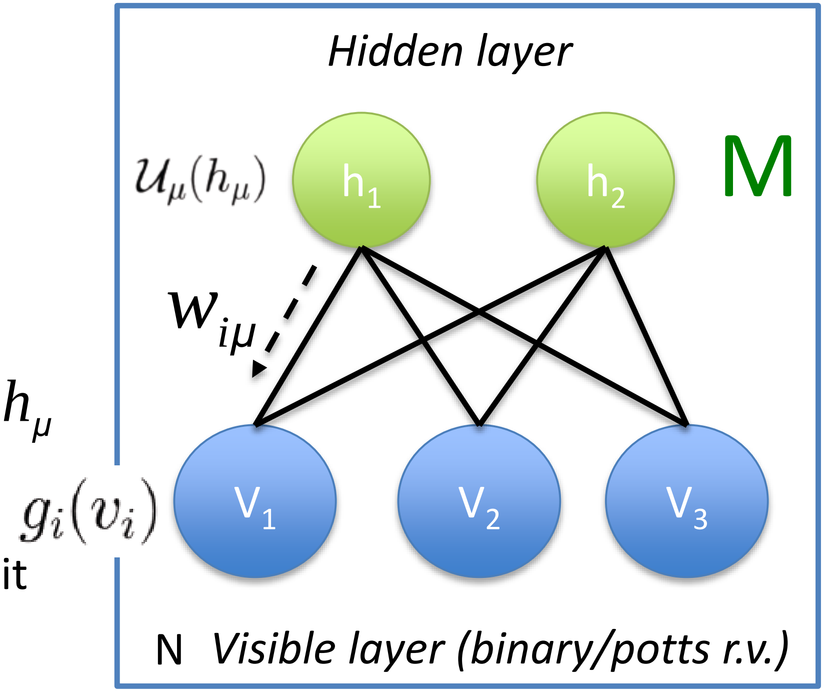
$$I_\mu(\mathbf{v}) = \sum_i w_{i\mu}(v_i)\ .$$

Which determines the probability of its activity:

$$P(h_\mu|\mathbf{v}) \propto \exp\left(-\mathcal{U}_\mu(h_\mu) + h_\mu\, I_\mu(\mathbf{v})\right)$$

Restricted Boltzmann Machines

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



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

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

.Given an hidden unit configuration the visible unit takes the value v_i with probability

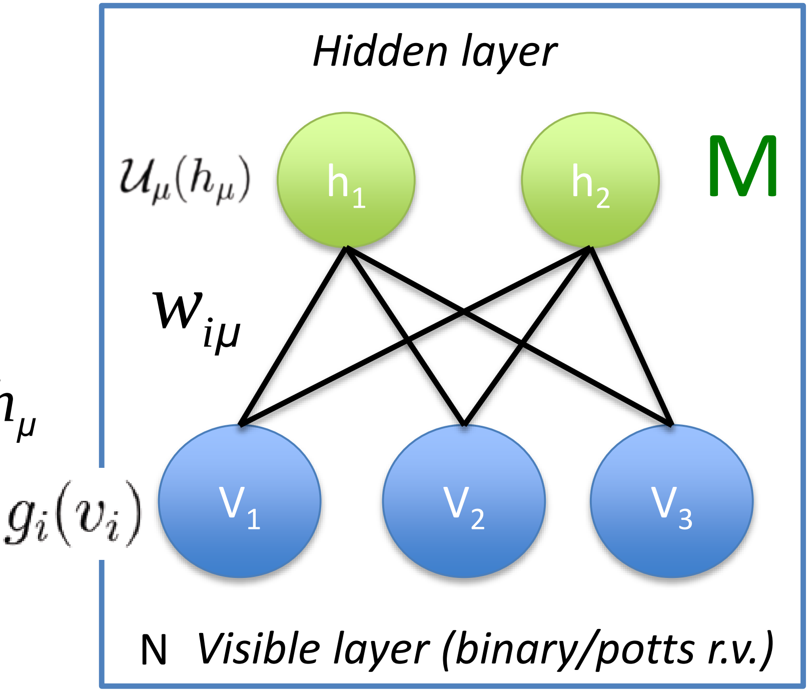
$$P(v_i|\mathbf{h}) \propto \exp\left(g_i(v_i) + \sum_\mu h_\mu w_{i\mu}(v_i)\right) .$$

Restricted Boltzmann Machines

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

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

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



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

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

RBM are generative models, trained through unsupervised learning

Learning is done by finding parameters maximizing the Likelihood

Parameters of RBM and data-representational phases

- Number of Hidden Units M
- Shape and parameters of Potential $\mathcal{U}_\mu(h_\mu)$
- Input Fields $g_i(v_i)$ and weights $w_{i\mu}$

and their sparsity (by adding a L_1 regularization)

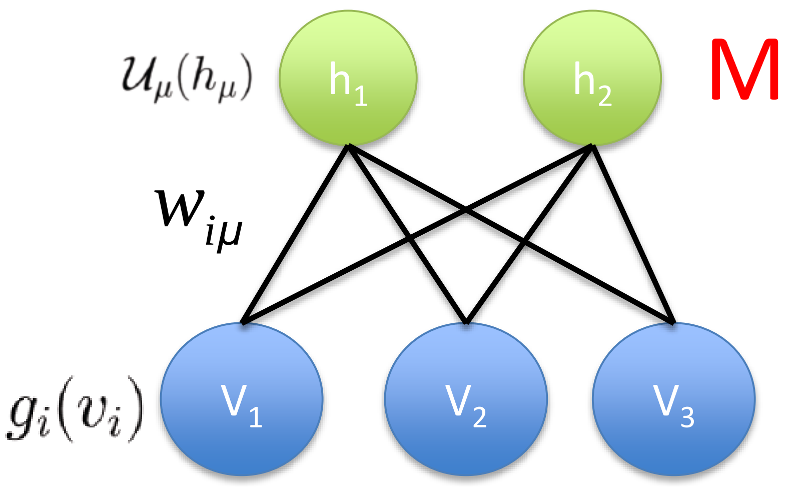
Parameters determined through training

Depending on such parameters there are different data-representational phases,separated by phase transitions

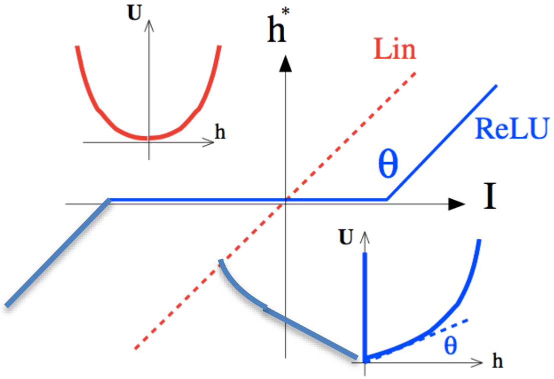
[J. Tubiana, R.Monasson, Physical Review Letters 118, 138301 (2017)]

We use Double Relu Units

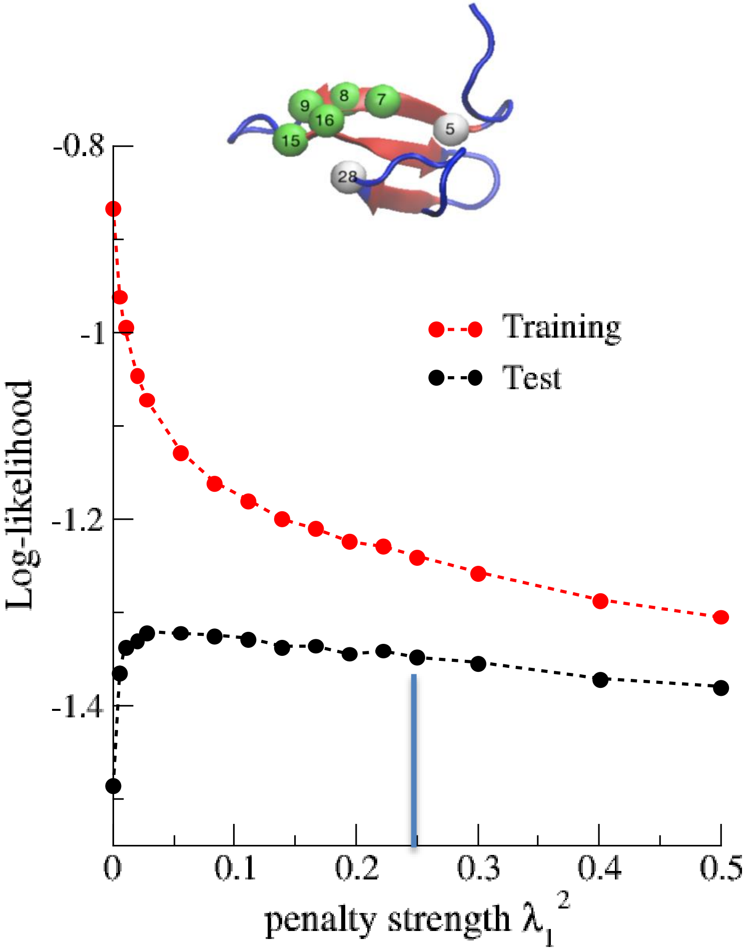
Hidden layer



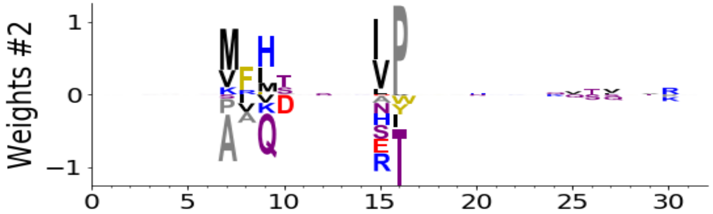
N *Visible layer (binary/potts r.v.)*



The interpretability-performance trade-off

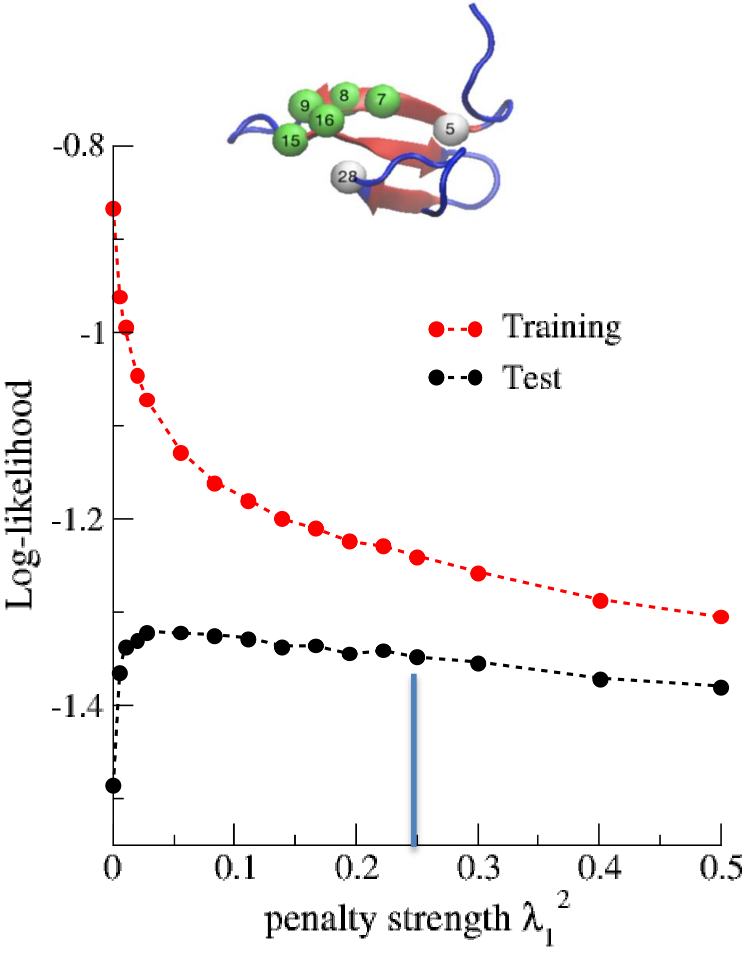


$/_1^2 = 0.25$

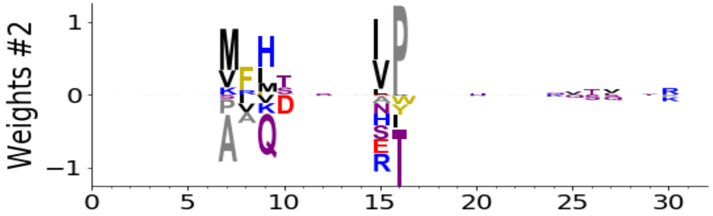


$$\langle \log P(\mathbf{v}) \rangle_{MSA} - \frac{\lambda_f}{2} \sum_{i,v} g_i(v)^2 - \frac{\lambda_1^2}{2qN} \sum_{\mu} \left(\sum_{i,v} |w_{i\mu}(v)| \right)^2 ,$$

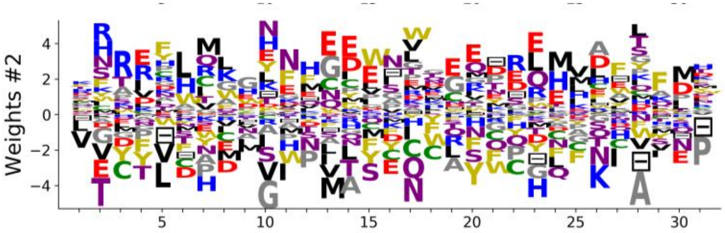
The interpretability-performance trade-off



$\lambda_1^2 = 0.25$

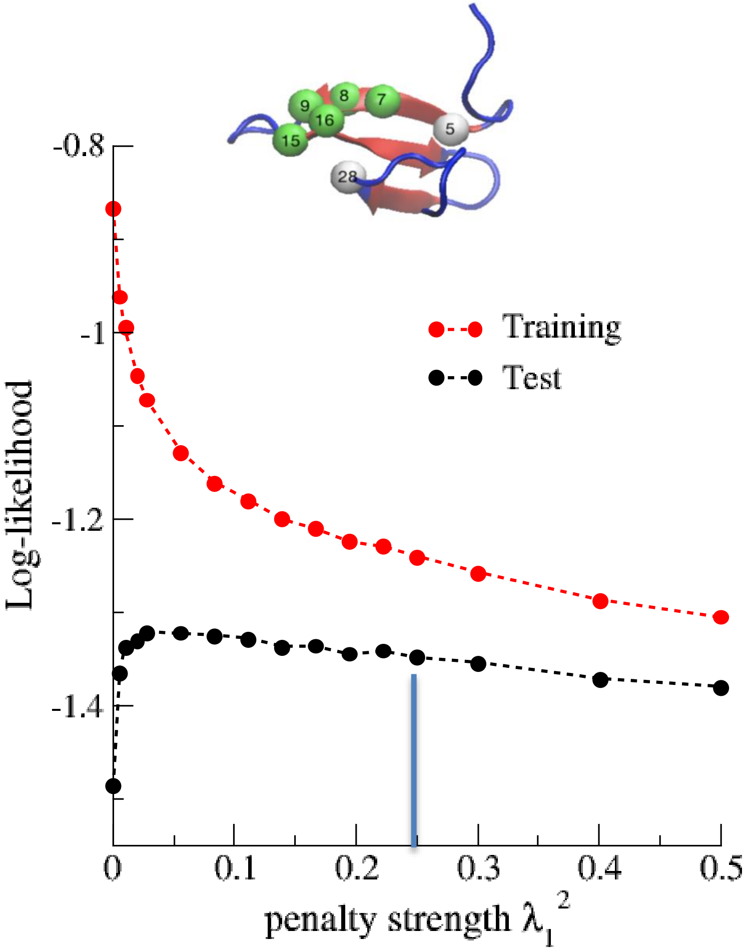


$\lambda_1^2 = 0$

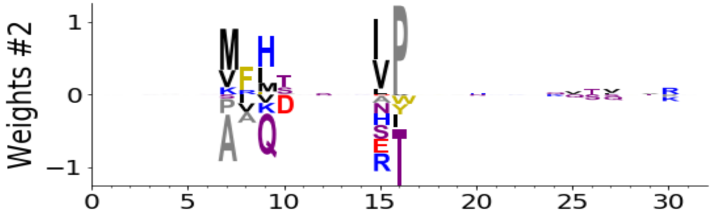


$$\langle \log P(\mathbf{v}) \rangle_{MSA} - \frac{\lambda_f}{2} \sum_{i,v} g_i(v)^2 - \frac{\lambda_1^2}{2qN} \sum_{\mu} \left(\sum_{i,v} |w_{i\mu}(v)| \right)^2$$

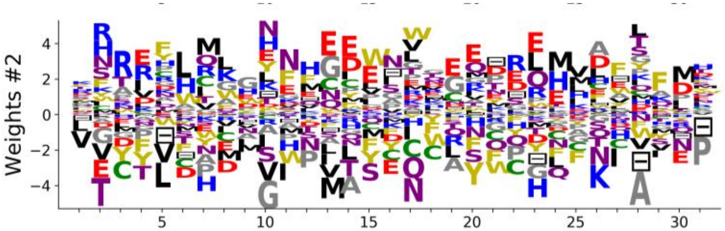
The interpretability-performance trade-off



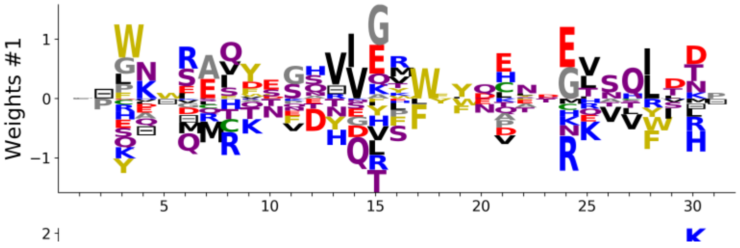
$/_{1}^{2} = 0.25$



$/_{1}^{2} = 0$



$/_{1}^{2} = 0.03$



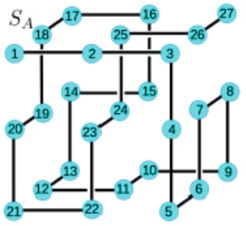
$$\langle \log P(\mathbf{v})\rangle_{MSA} - \frac{\lambda_f}{2}\sum_{i,v} g_i(v)^2 - \frac{\lambda_1^2}{2qN}\sum_{\mu}\left(\sum_{i,v}|w_{i\mu}(v)|\right)^2 ,$$

Tubiana Cocco Monasson , elife, 2019,

Protein families studied

- Lattice Proteins

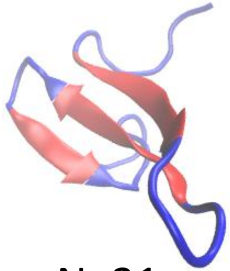
Shaknovich et al. J. Chem. Phys. 1990
Jacquin et al. PLOS CB 2016



N=27 aa

- WW Domain

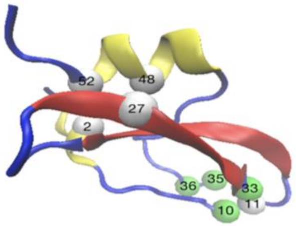
Russ et al. Nature 2005



N=31 aa

- Kunitz Domain

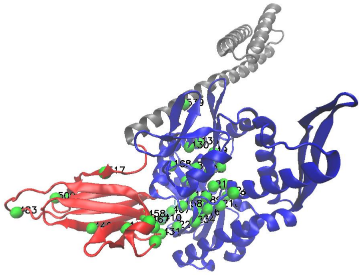
Morcos et al. PNAS 2011



N=54 aa

- Hsp70 chaperone

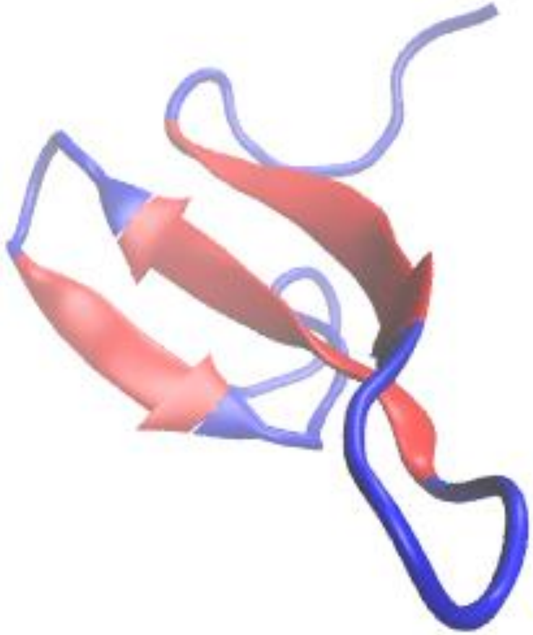
Smock et al. Mol. Sys. Biol. 2010
Malinverni et al. PLOS CB 2015



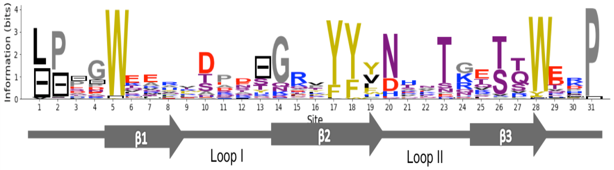
N=661 aa

The WW Domain

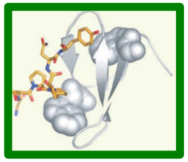
- N=30-40 amino-acids (very small)
- Role:
  - Gene regulation, transcription
  - RNA processing
  - Receptor signaling
- Recognition of Proline-Rich Linear Motifs
- 4 types of ligand specificities



WW: a small binding domain

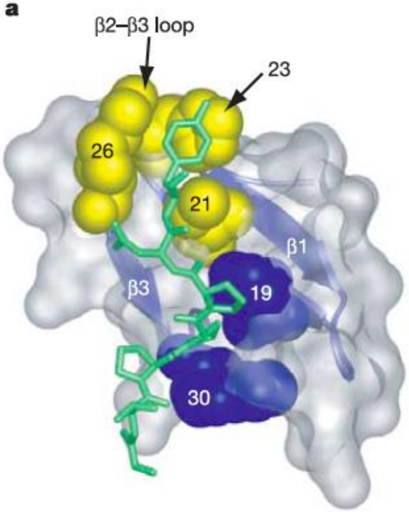


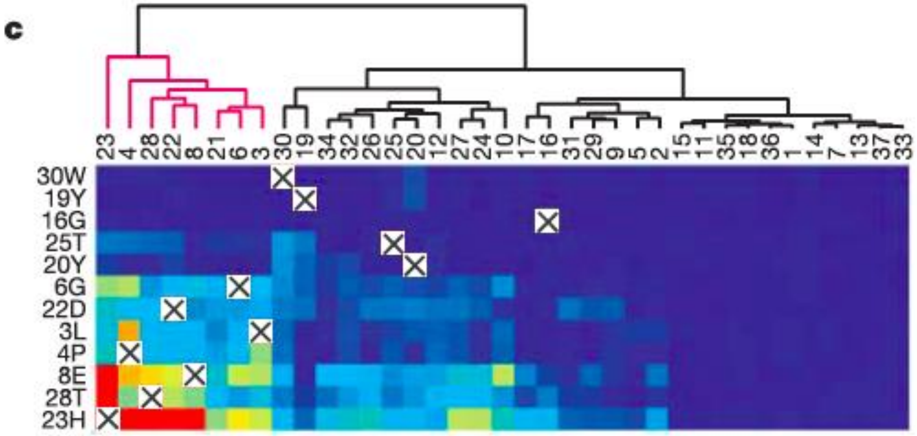
```
● PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
● PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
● PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
● LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
● GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
● GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
● -MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
● SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
● DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
● AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```



Sequences have different binding affinity:

Sequence	Ligand
group I:	PPxY
group II:	PPLP
group III:	PPR
group IV:	PS/PT

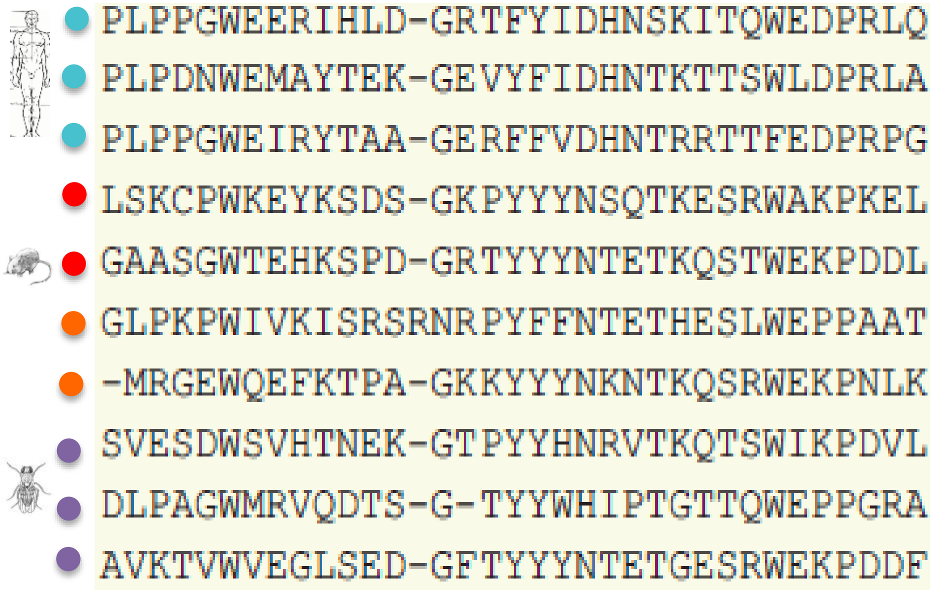




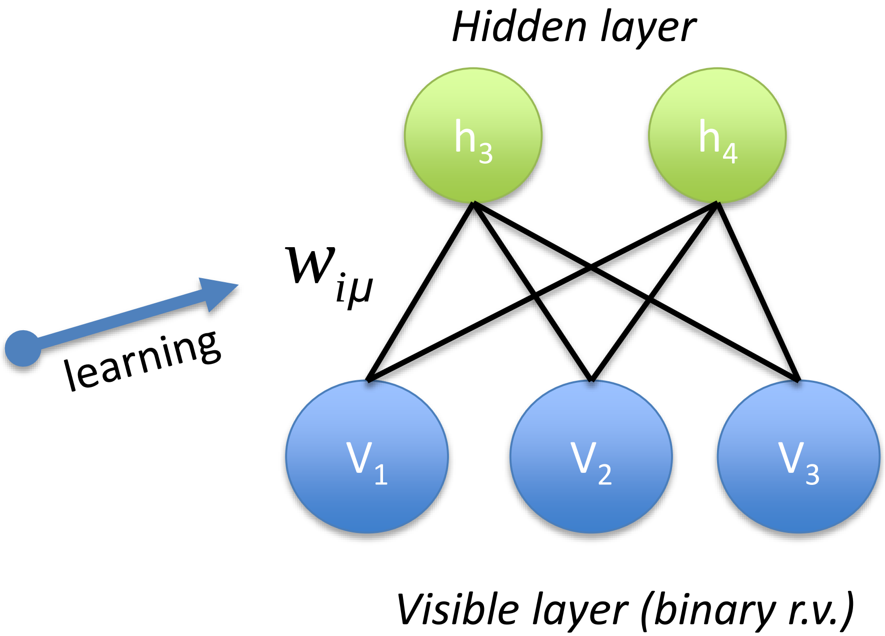
Sector Analysis: 8 positions very correlated

[W.P. Russ…R. Ranganathan, Nature 2005
N Halabi,…R.Ranganathan 2009]

RBM features

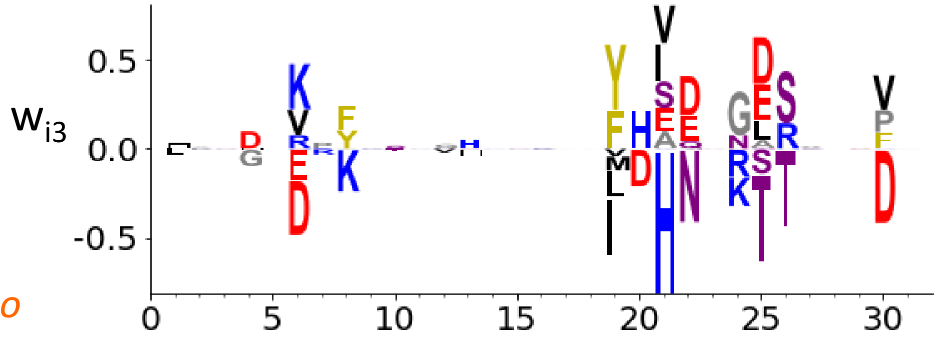


```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```



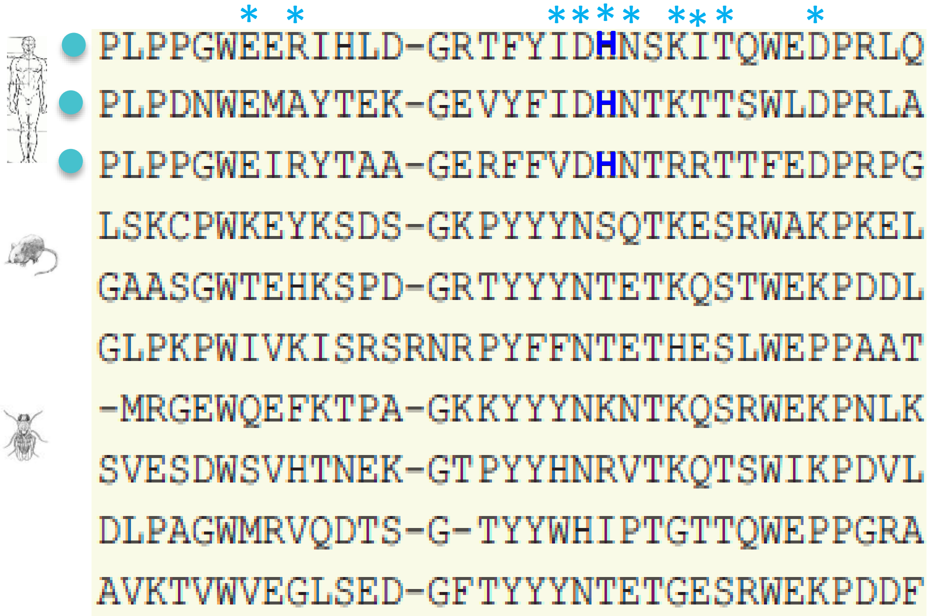
Similarly to principal components in PCA: features $w_{i\mu}$

[A Raussel.. A Valencia 2010, N Halabi,…R.Ranganathan 2009]

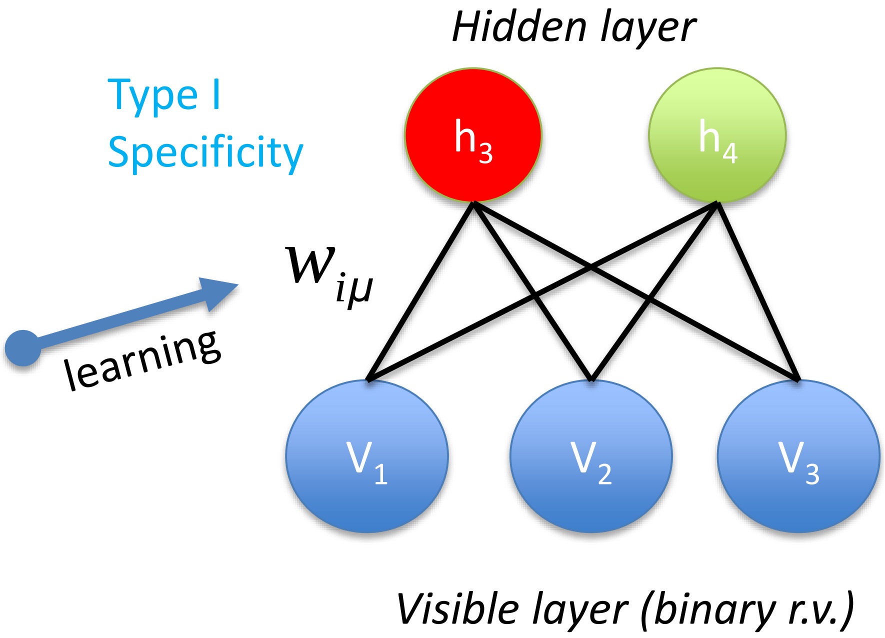


Tubiana Cocco Monasson 2018, arXiv: 1803.08718 q.bio

RBM features

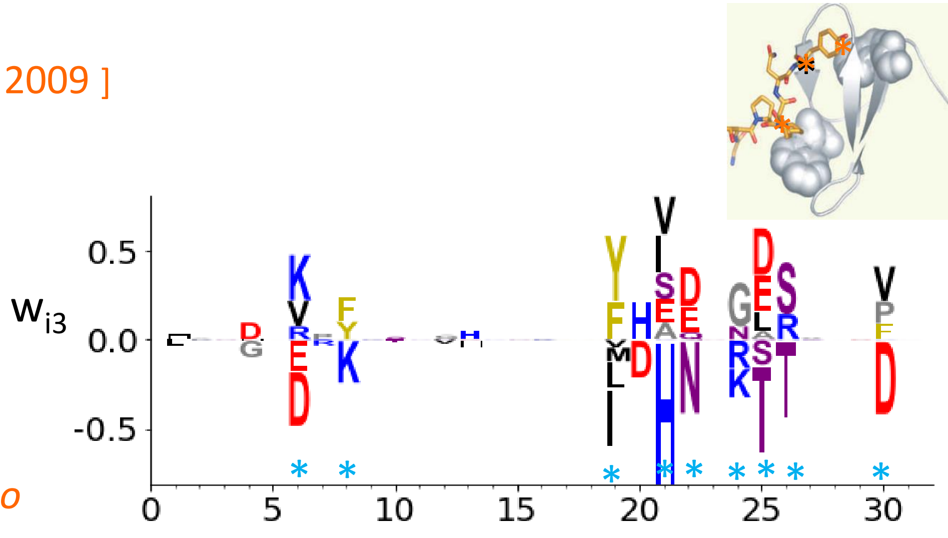


```
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```



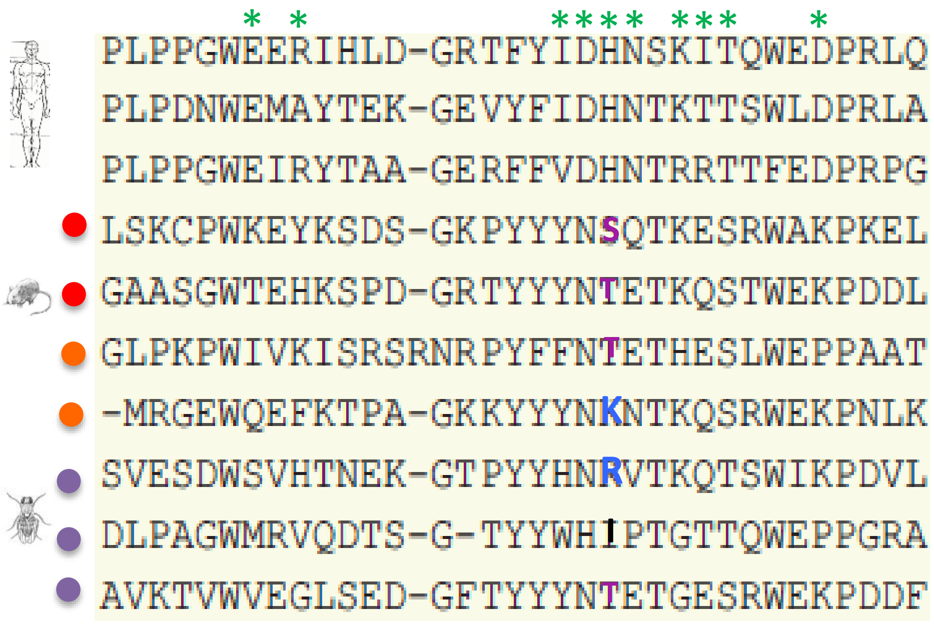
Similarly to principal components in PCA: features $w_{i\mu}$

[A Raussel.. A Valencia 2010, N Halabi,…R.Ranganathan 2009]



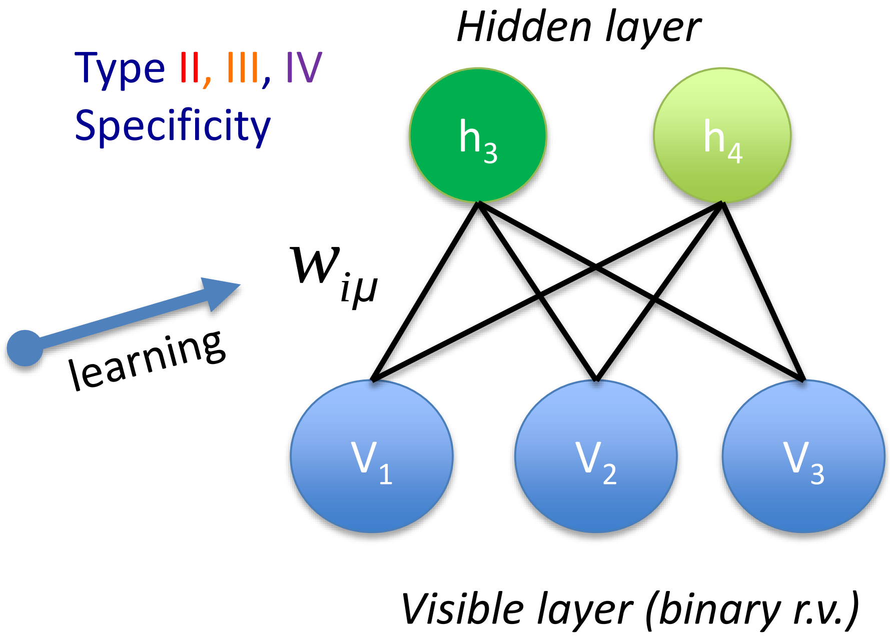
Tubiana Cocco Monasson 2018, arXiv: 1803.08718 q.bio

RBM features



```
      * *          **** ***   *
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
PLPDNWEMAYTEK-GEVYFIDHNTKTTSWLDPRLA
PLPPGWEIRYTAA-GERFFVDHNTRRTTFEDPRPG
LSKCPWKEYKSDS-GKPYYYNSQTKESRWAKPKEL
GAASGWTEHKSPD-GRTYYYNTETKQSTWEKPDDL
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT
-MRGEWQEFKTPA-GKKYYYNKNTKQSRWEKPNLK
SVESDWSVHTNEK-GTPYYHNRVTKQTSWIKPDVL
DLPAGWMRVQDTS-G-TYYWHIPTGTTQWEPPGRA
AVKTVWVEGLSED-GFTYYYNTETGESRWEKPDDF
```

Type II, III, IV
Specificity

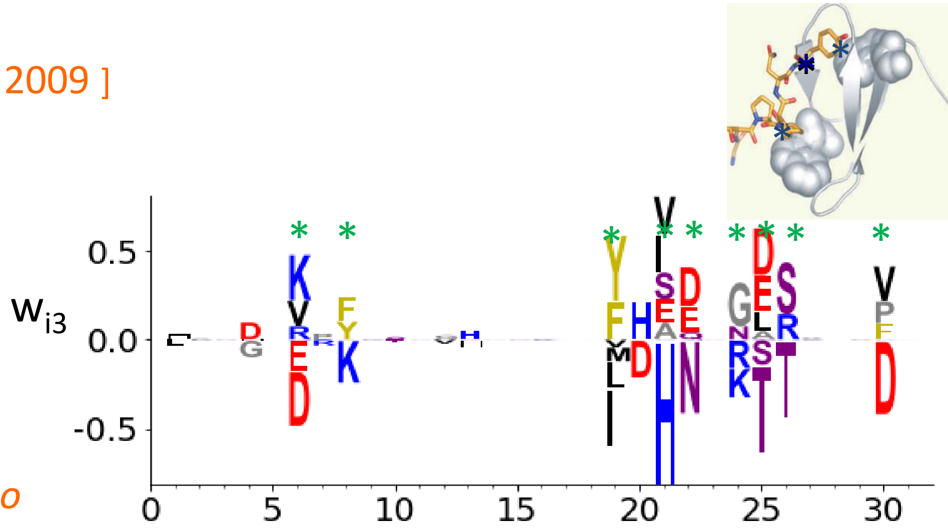


Hidden layer

Visible layer (binary r.v.)

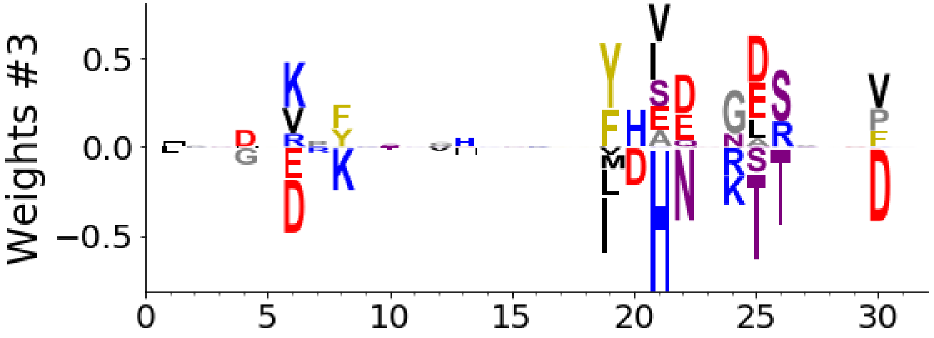
Similarly to principal components in PCA: features $w_{i\mu}$

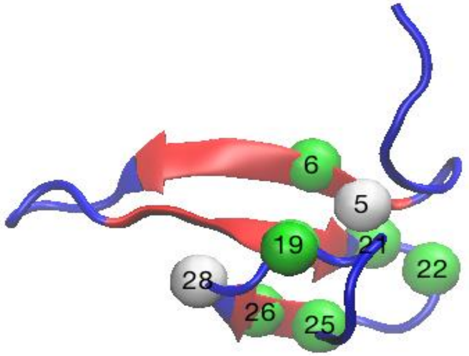
[A Raussel.. A Valencia 2010, N Halabi,…R.Ranganathan 2009]

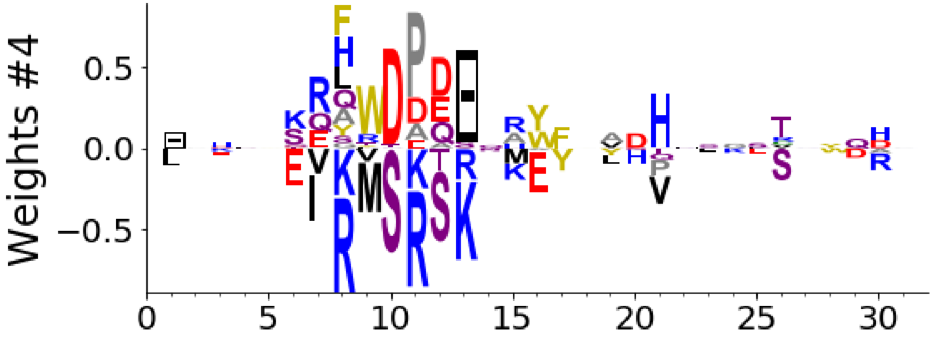


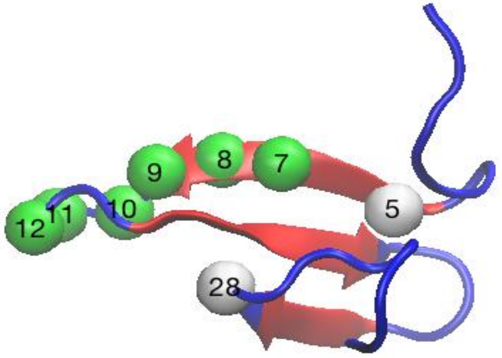
Tubiana Cocco Monasson 2018, arXiv: 1803.08718 q.bio

RBM features

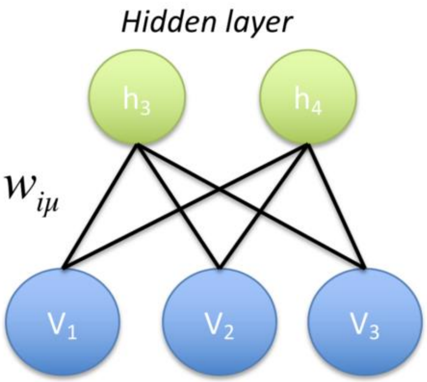






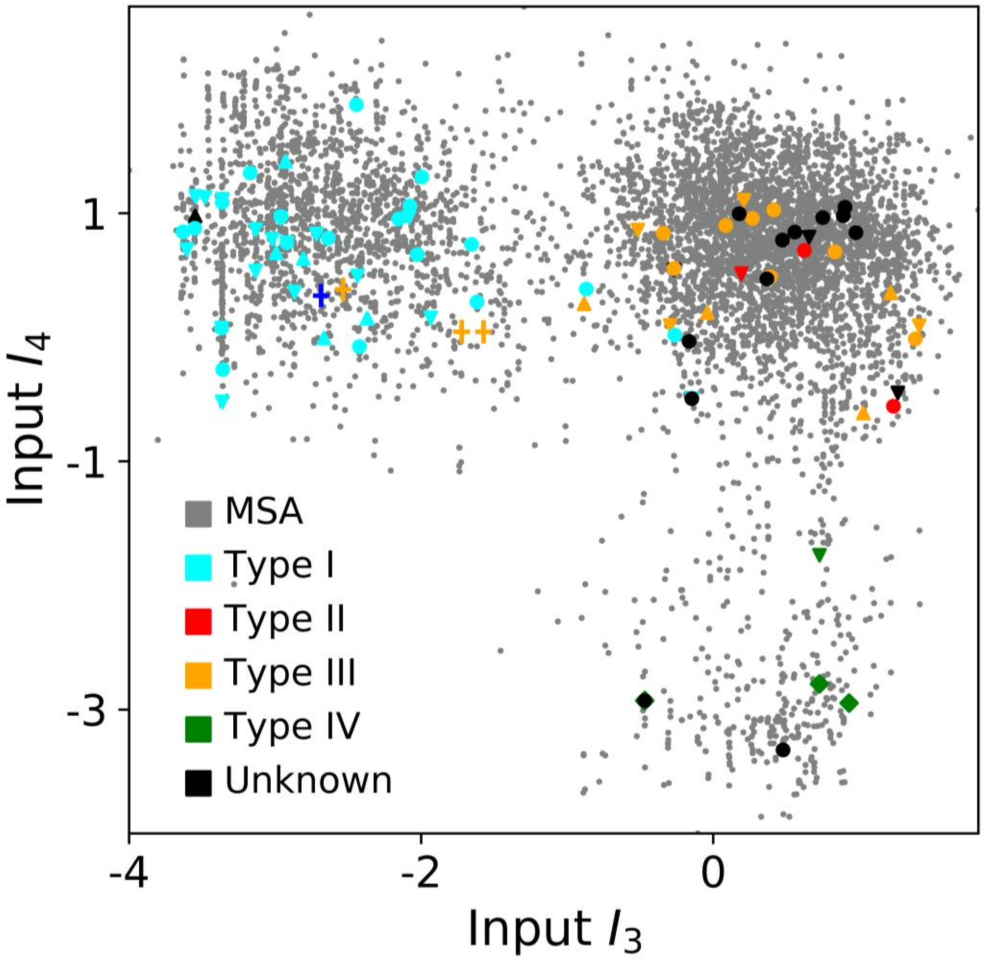


Tubiana Cocco Monasson , elife, 2019,



RBM features reflect specificity

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



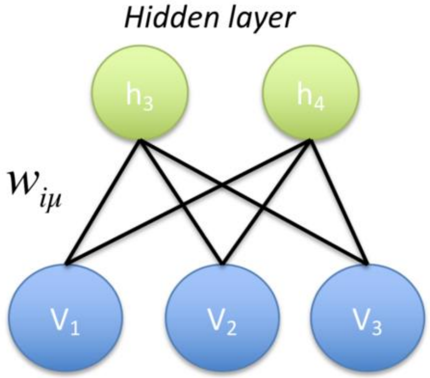
Experimental data from:

Russ et al. Nature 2005

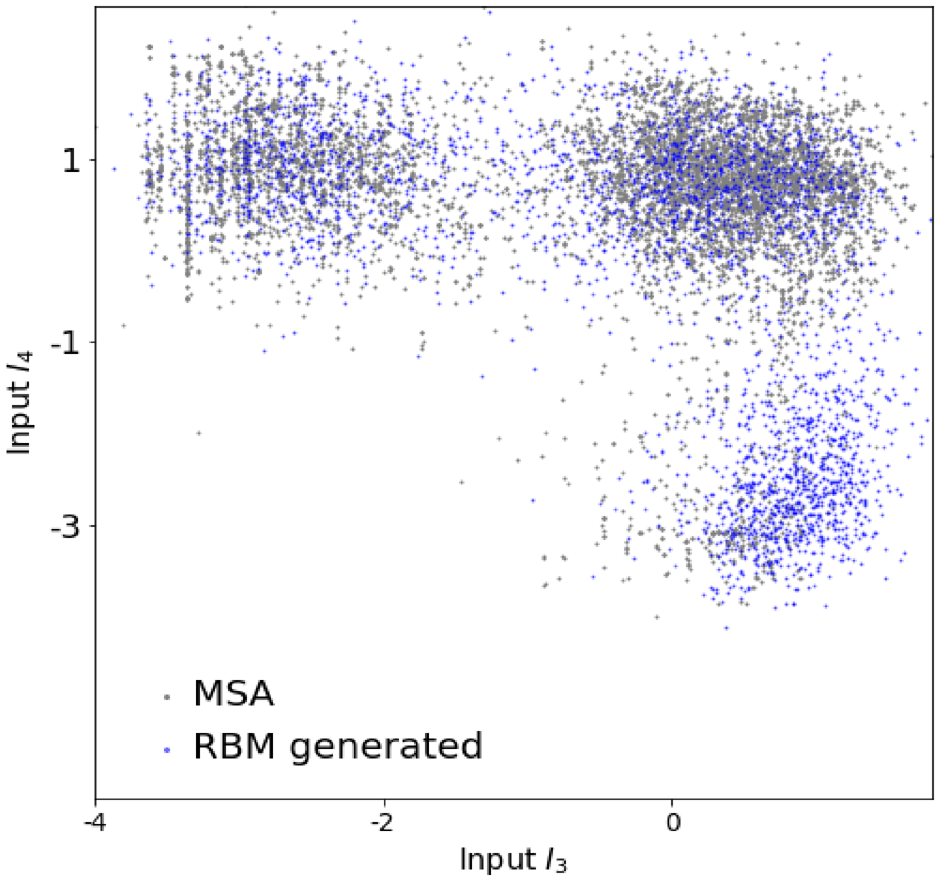
Espanel and Sudol J. Biol. Chem. 1999

Otte et al. Protein Science 2003

Tubiana Cocco Monasson , elife, 2019,



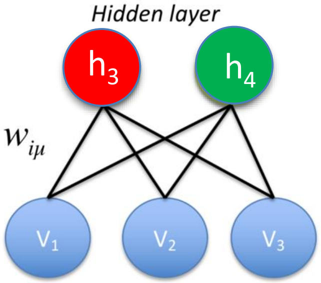
Artificial Sequence Generation with RBM

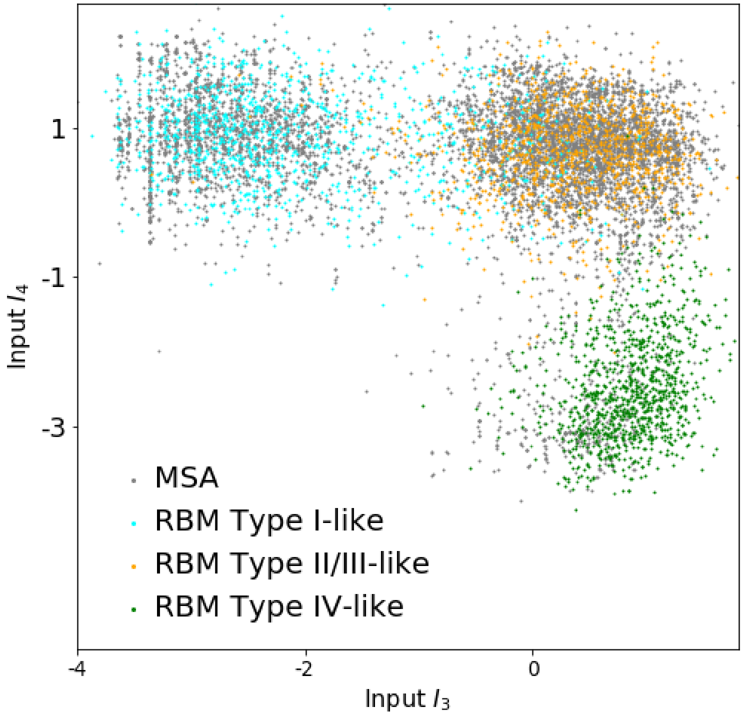


Tubiana Cocco Monasson , elife, 2019,

Artificial Sequence Generation with RBM

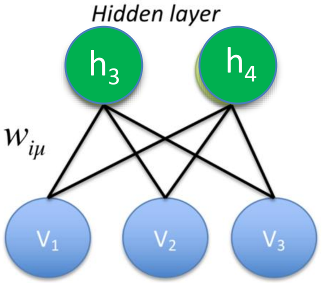
Type I-like
binding specificity +
Short loop



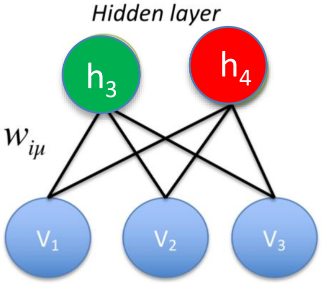


Artificial Sequences

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

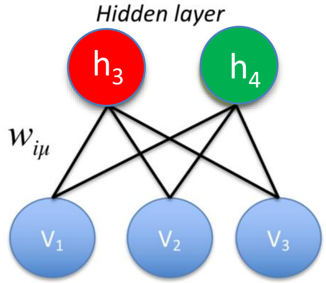


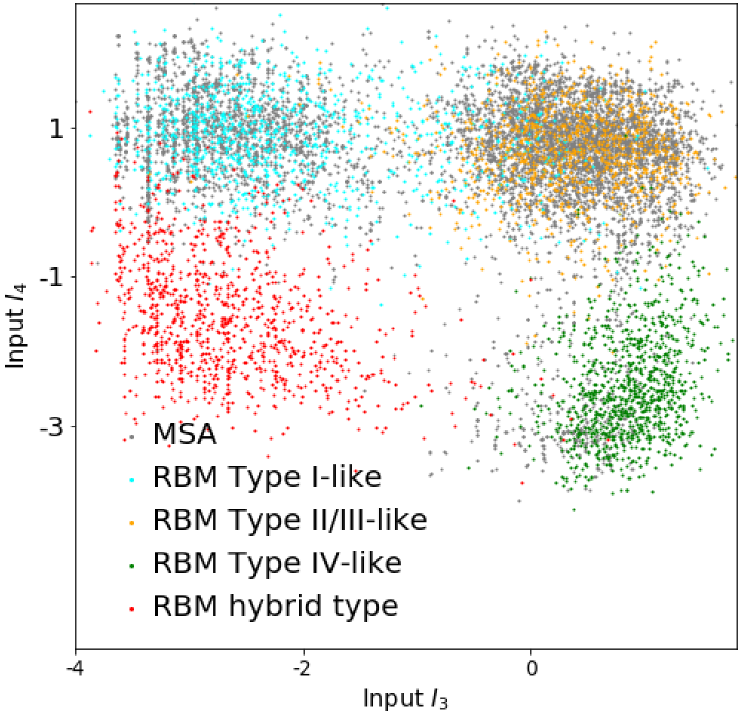
Type II/III/IV-like
binding pocket +
Long loop → Type IV



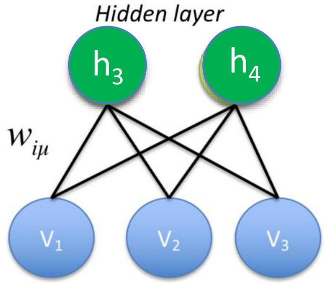
Artificial Sequence Generation with RBM

Type I-like
binding specificity +
Short loop

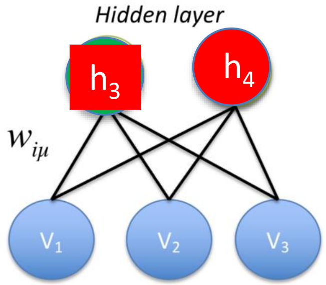




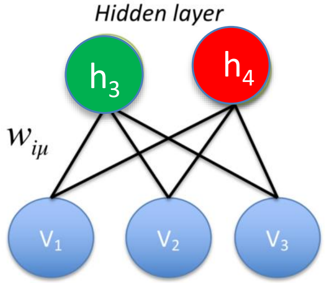
Type II/III/IV-like
binding specificity +
Short loop → Type II/III



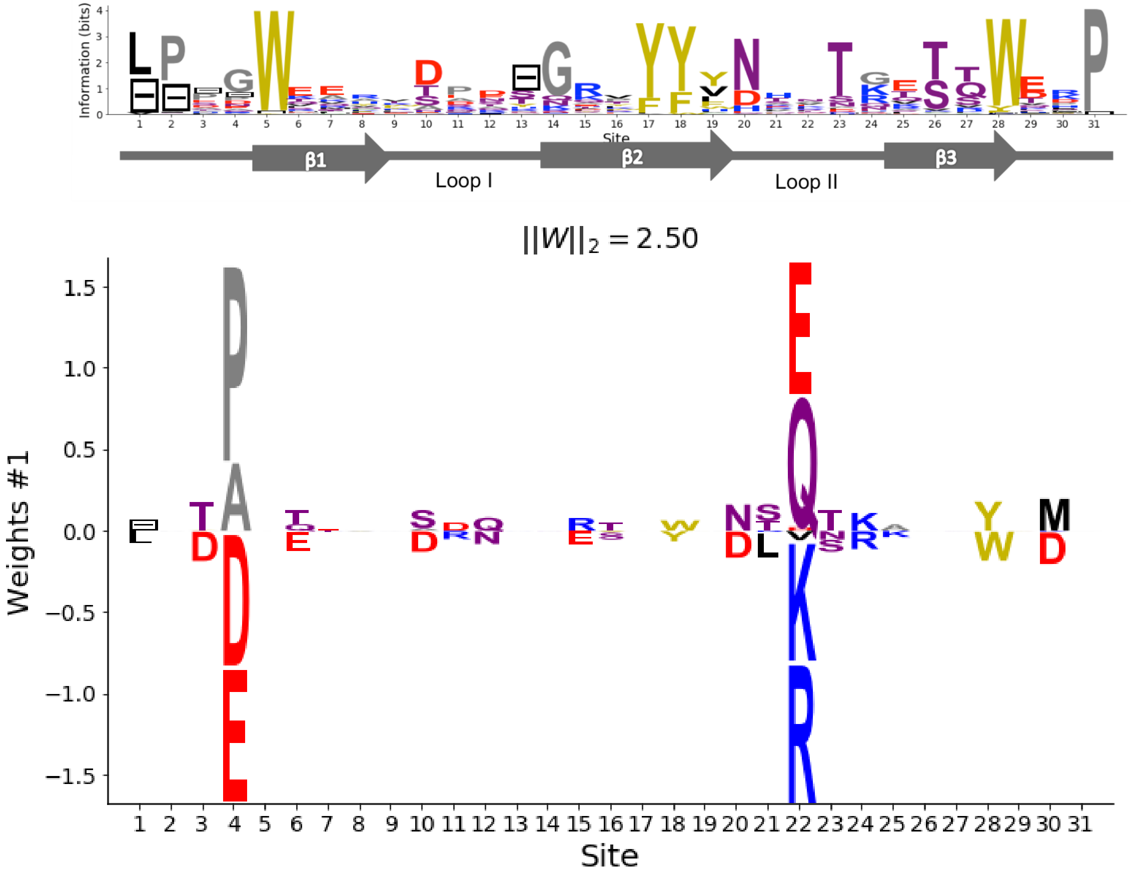
Type II/III/IV-like
binding pocket +
Long loop → Type IV



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



RBM WW Features: A contact mode



Hsp70 chaperone protein

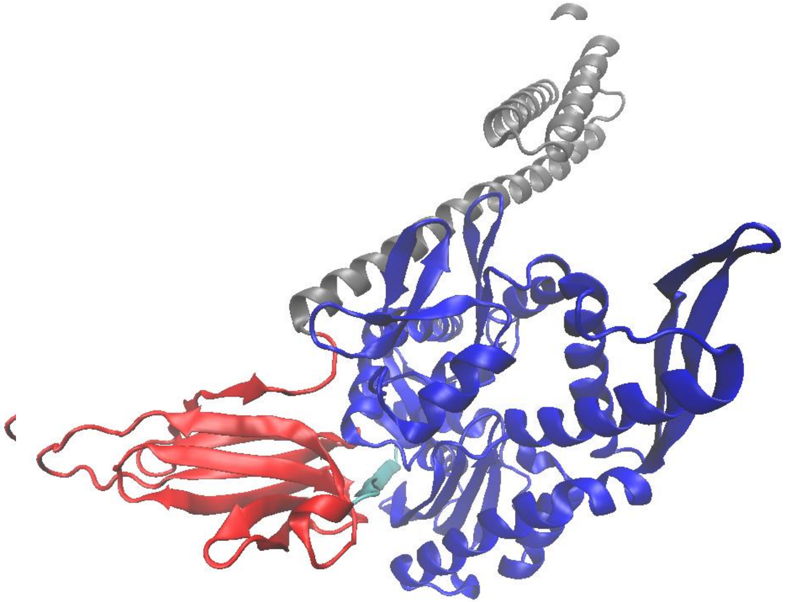
- N>600 amino-acids
- Multidomain.
  - Nucleotide Binding Domain (NBD)
  - Substrate Binding Domain (SBD)
  - LID Domain
  - Linker

Function:

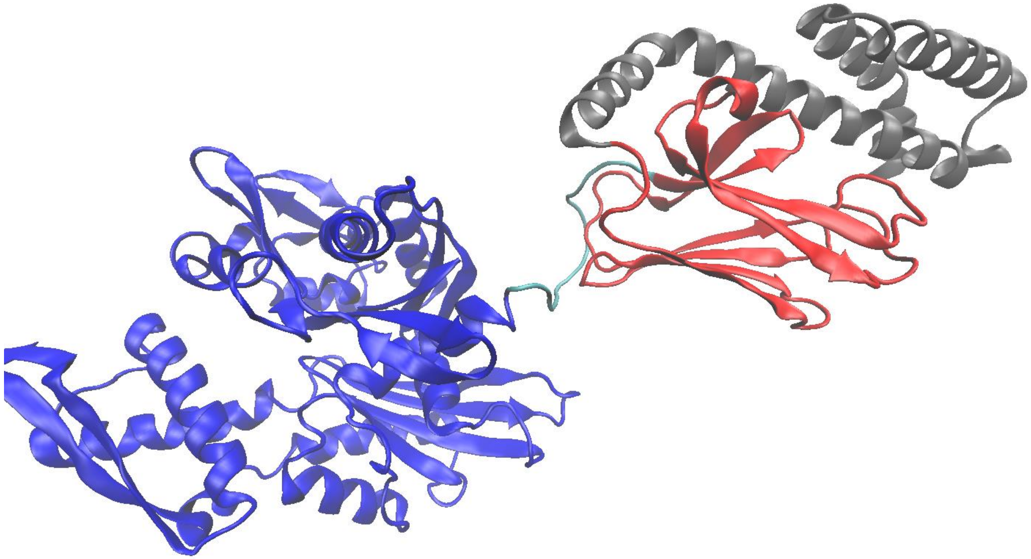
- Traps substrate proteins between the LID and the SBD
- LID/SBD cavity is either open or close

Roles:

- Assist protein folding
- Transport proteins for degradation

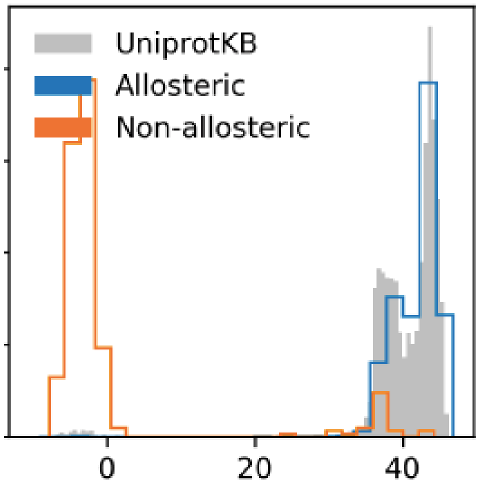


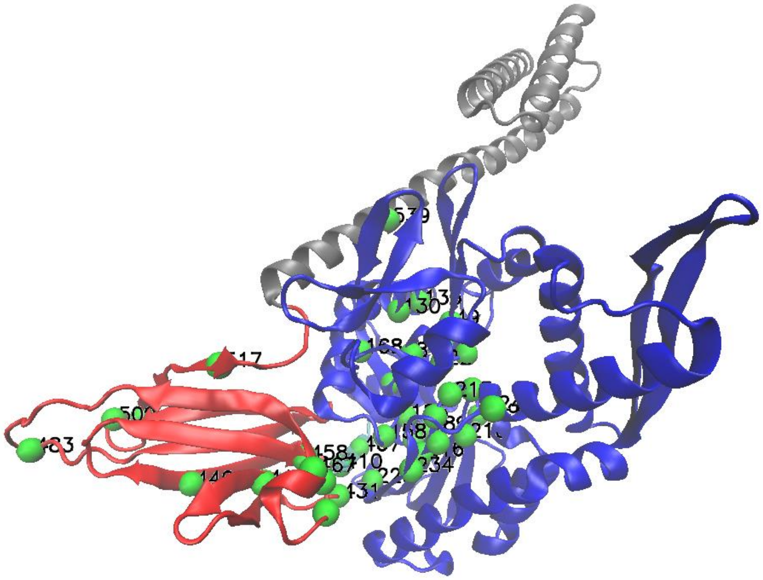
ATP bound conformation (open)

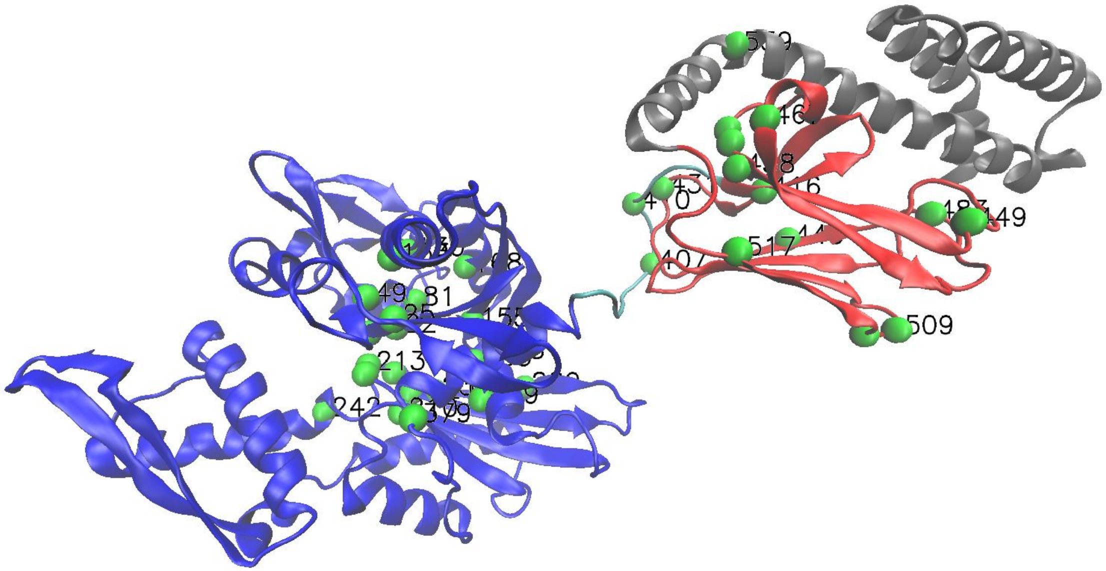


ADP bound conformation (closed)

Interdomain features control allostery







Conclusion

- Summary:
  - Under specific conditions (weight sparsity, non-linearity), RBM learn compositional representations of data.
  - They achieve a good trade-off between interpretability and performance
  - RBM can extract meaningful features from sequence

  and cluster protein subfamilies with respect to different properties eg. stability, binding specificity, allostery..

  - RBM can Generate sequences with specific properties (in given clusters)

- But:

-RBM less well known and studied Model than BM. Training not guarantee to work well: Log-Likelihood is not a convex function …

- Outlook:
  - Experimental validation of designed sequences