How pairwise coevolutionary models capture the collective residue variability in proteins

Matteo Figliuzzi, Pierre Barrat-Charlaix, Martin Weigt

UPMC SORBONNE UNIVERSITÉS

LCQB

Statistical modeling of protein sequences

Protein family

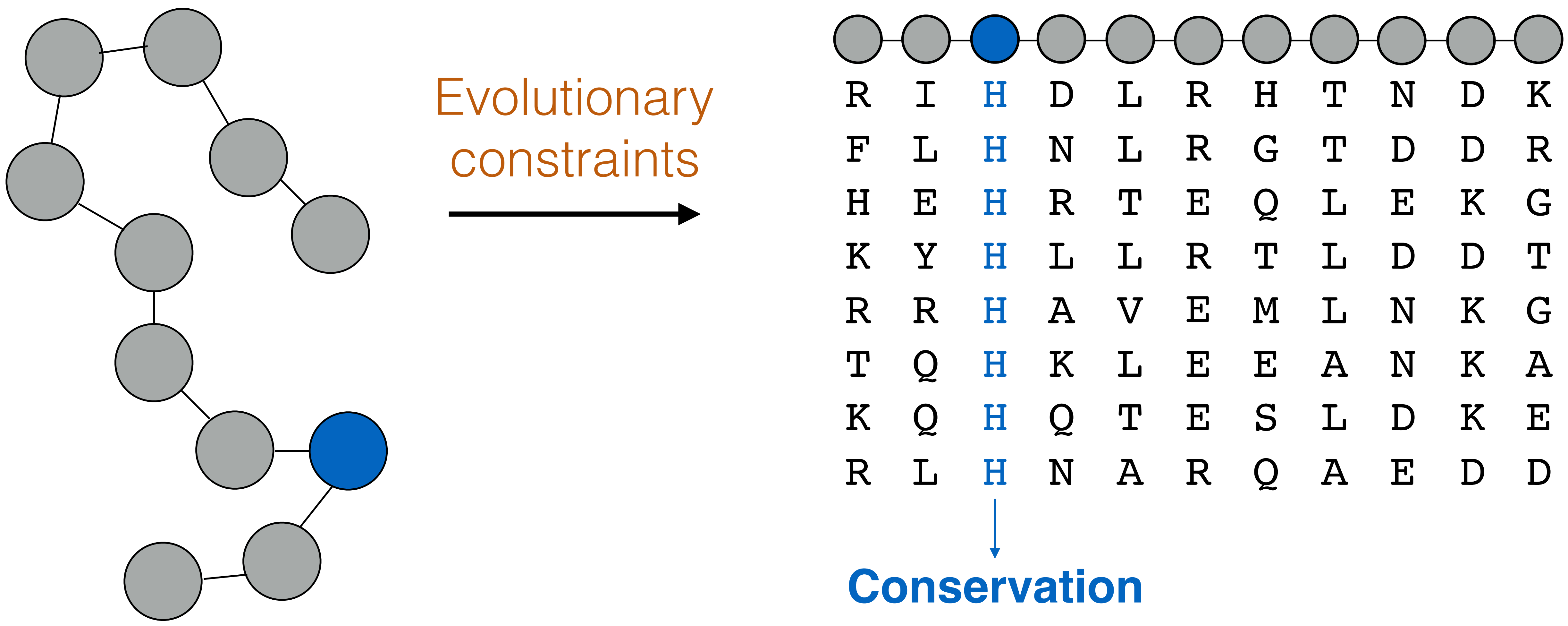
Evolutionary constraints →

Multiple Sequence Alignment

```
...
YHCDKCSMSFAAPSRLNKHMRTH
HKCSYCSKAFIKKTLLKAHERTH
-QCEECGKQFAYSHSLKTHMMTH
YVCNVCGNLFRQHSTLTIHMRTH
-TCEFCGKNFERNGNYVEHRRTH
FVCGVCNKGFNSRTYLLEHMNKH
YVCHFCGKAVTNRESLKTHVRLH
YSCNVCDKSFTQRSSLVVHQRTH
FECQICGKSFKRSVQLKYHMEIH
YKCATCQKSFKRSQELKSHGKLH
HACGICGKTFPNNSSLEKHKHIH
YVCDKCGRSFSQRSSLTIHQRYH
YTCNVCGKTVTTKKSYTNHVKIH
FKCGVCGKFYKNESSLKTHSKIH
-QCEECGEIFNHKSSLNKHLLKH
YACEYCDKRFGDKQYLTQHRRVH
FKCDECGQCFSQRSSLNRHKRYH
YECDICGICFNQRSTMTSHRRSH
...
```

Information?

Profile models



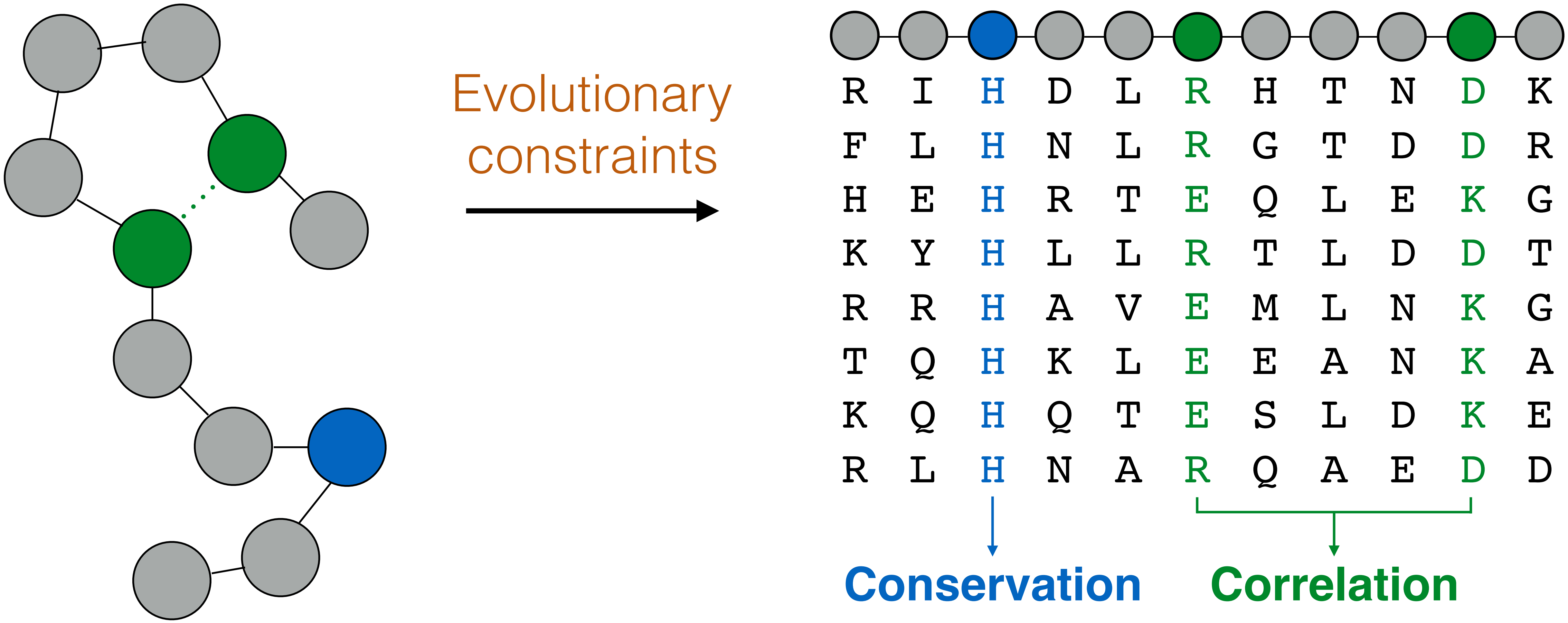
Evolutionary constraints

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

Conservation

- Functionally important **positions**
- Homology detection (HMM)
- **Unable to capture relations between columns**

Global statistical models



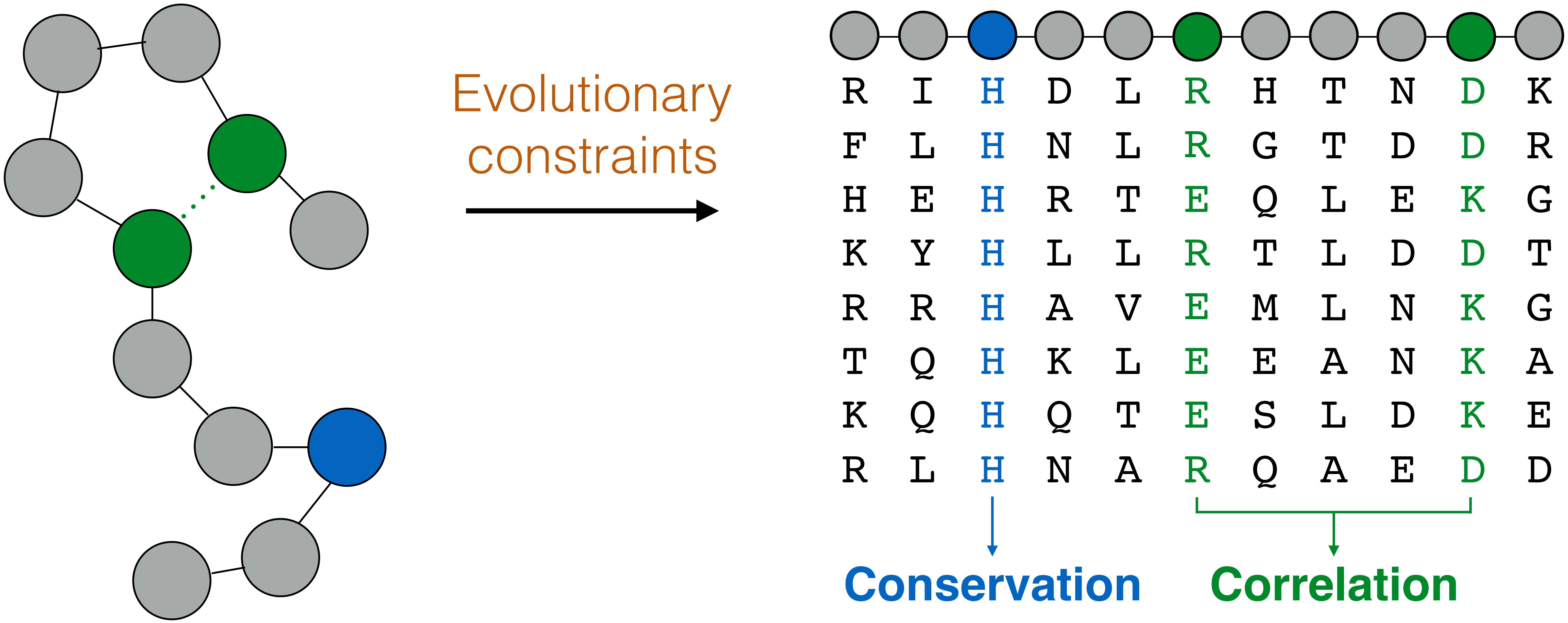
Evolutionary constraints

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

Conservation

Correlation

Global statistical models



Evolutionary constraints

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

Conservation

Correlation

$$P(a_1,\ldots,a_N)=\frac{1}{Z}\exp\left(\sum_{i,j=1}^{L}J_{ij}(a_i,a_j)+\sum_{i=1}^{L}h_i(a_i)\right)$$

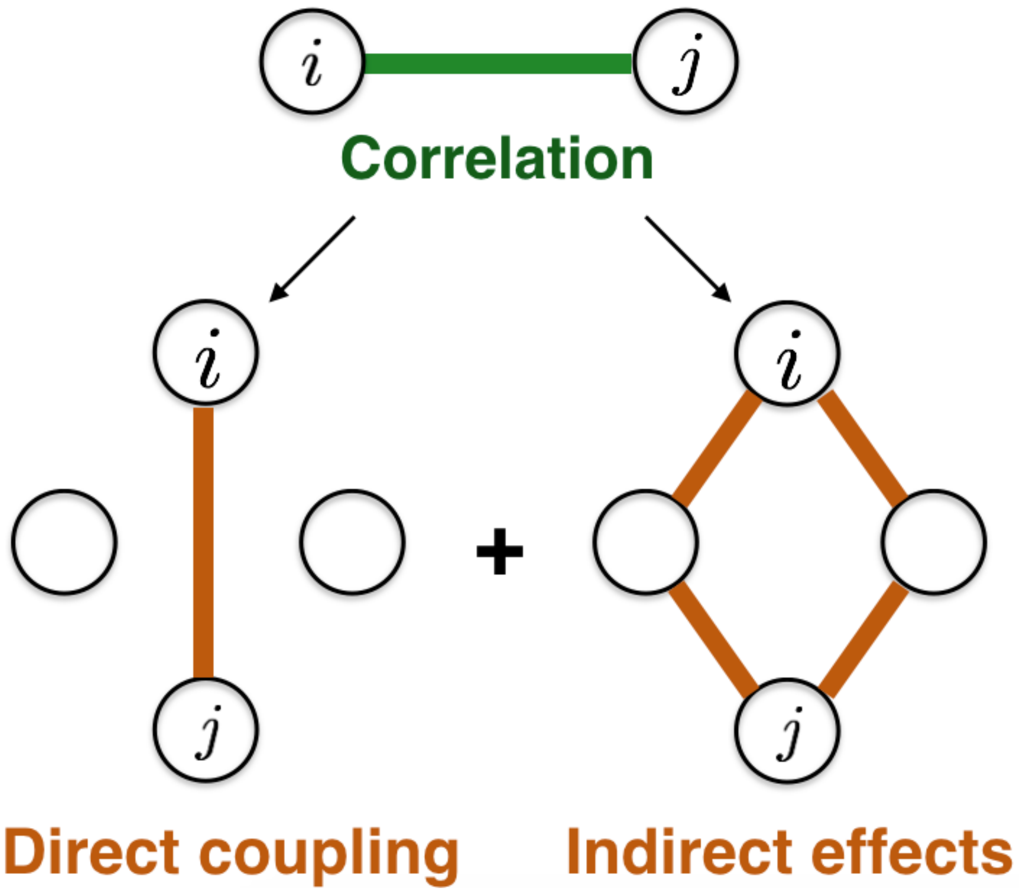
Direct Coupling Analysis (DCA)

- Intra/Inter protein **contacts**
- Protein-protein **interaction**
- Prediction of **mutational effects**
- **Generative model**

Global statistical models : the Potts model

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$

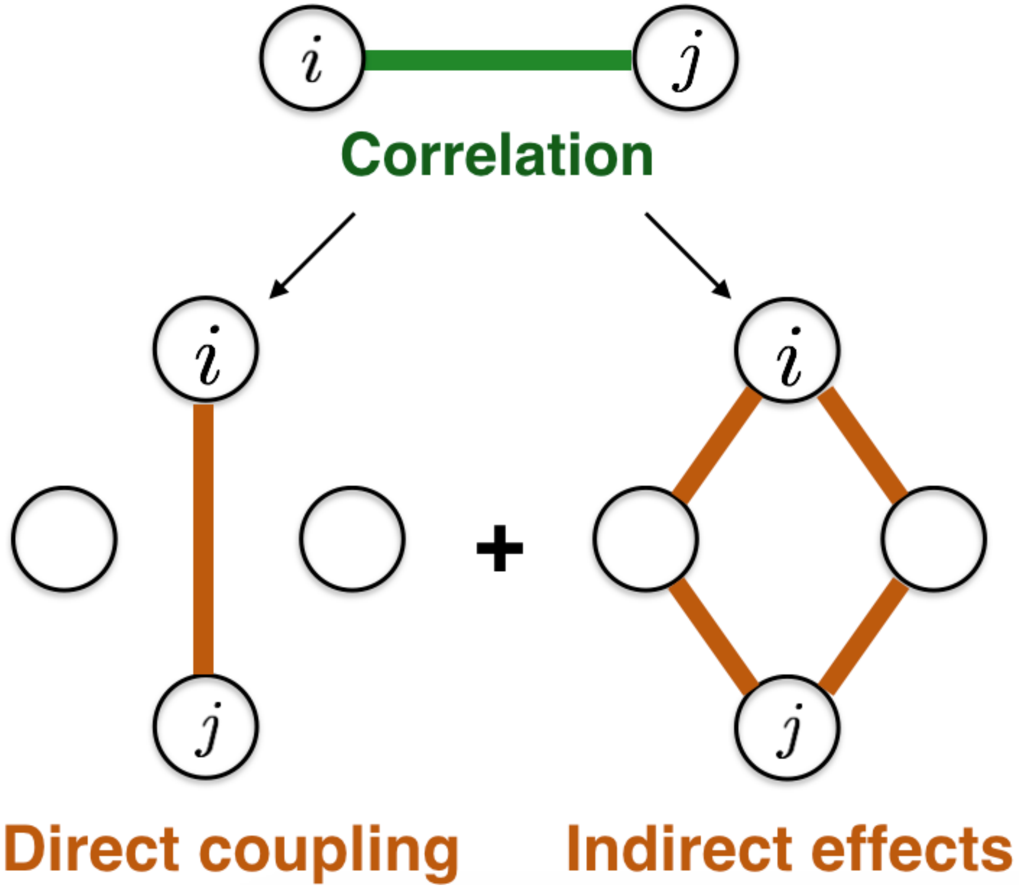
Disentangling correlations



Global statistical models : the Potts model

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp \left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$

Disentangling correlations



Maximum entropy modeling

Model with **maximal entropy** …

$$-\sum_{\{\vec{a}\}} P(\vec{a}) \log P(\vec{a}) \longrightarrow \text{Max}$$

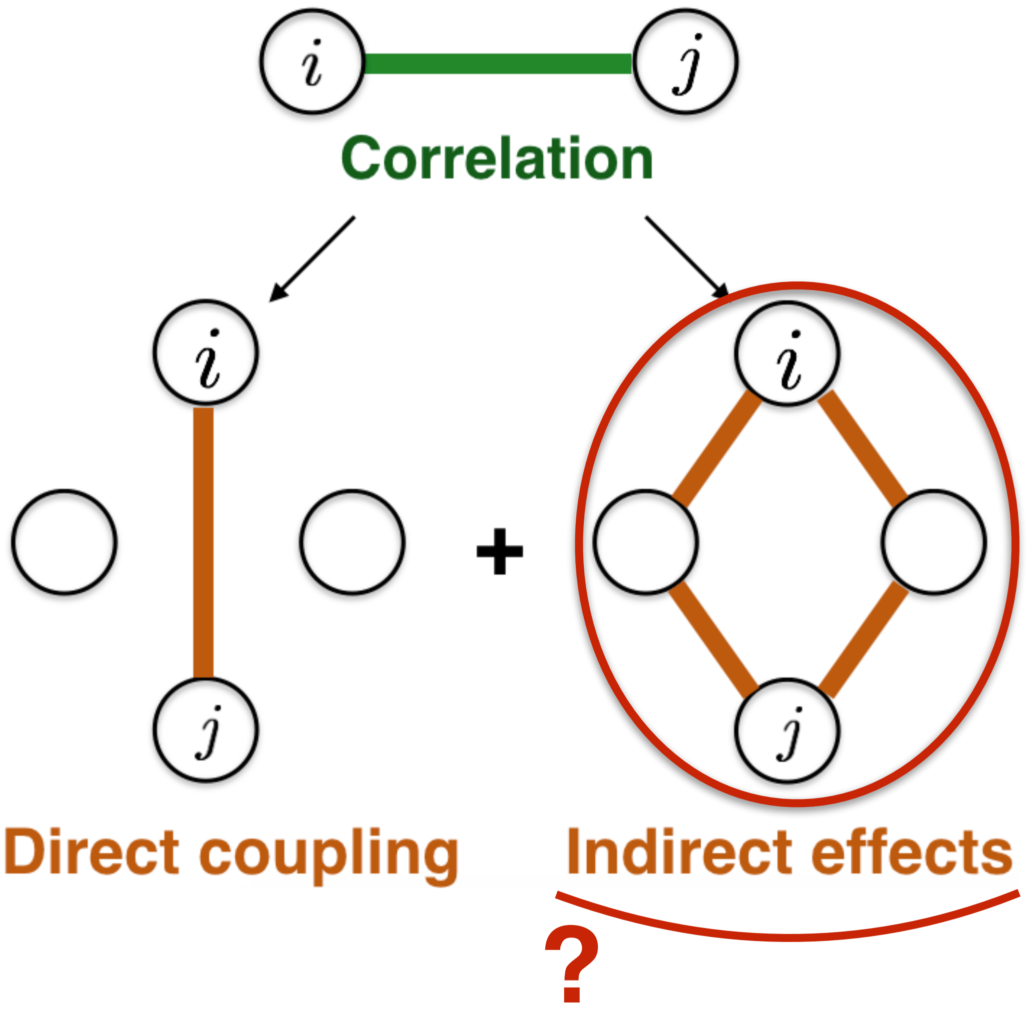
… while reproducing **pairwise statistics of data**

$$P_{ij}(a,b) = f_{ij}(a,b)$$

Global statistical models : the Potts model

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$

Disentangling correlations



Maximum entropy modeling

Model with **maximal entropy** …

$$-\sum_{\{\vec{a}\}} P(\vec{a}) \log P(\vec{a}) \longrightarrow \text{Max}$$

… while reproducing **pairwise statistics of data**

Why?

$$\cancel{P_{ij}(a,b) = f_{ij}(a,b)}$$

Inference based on approximations

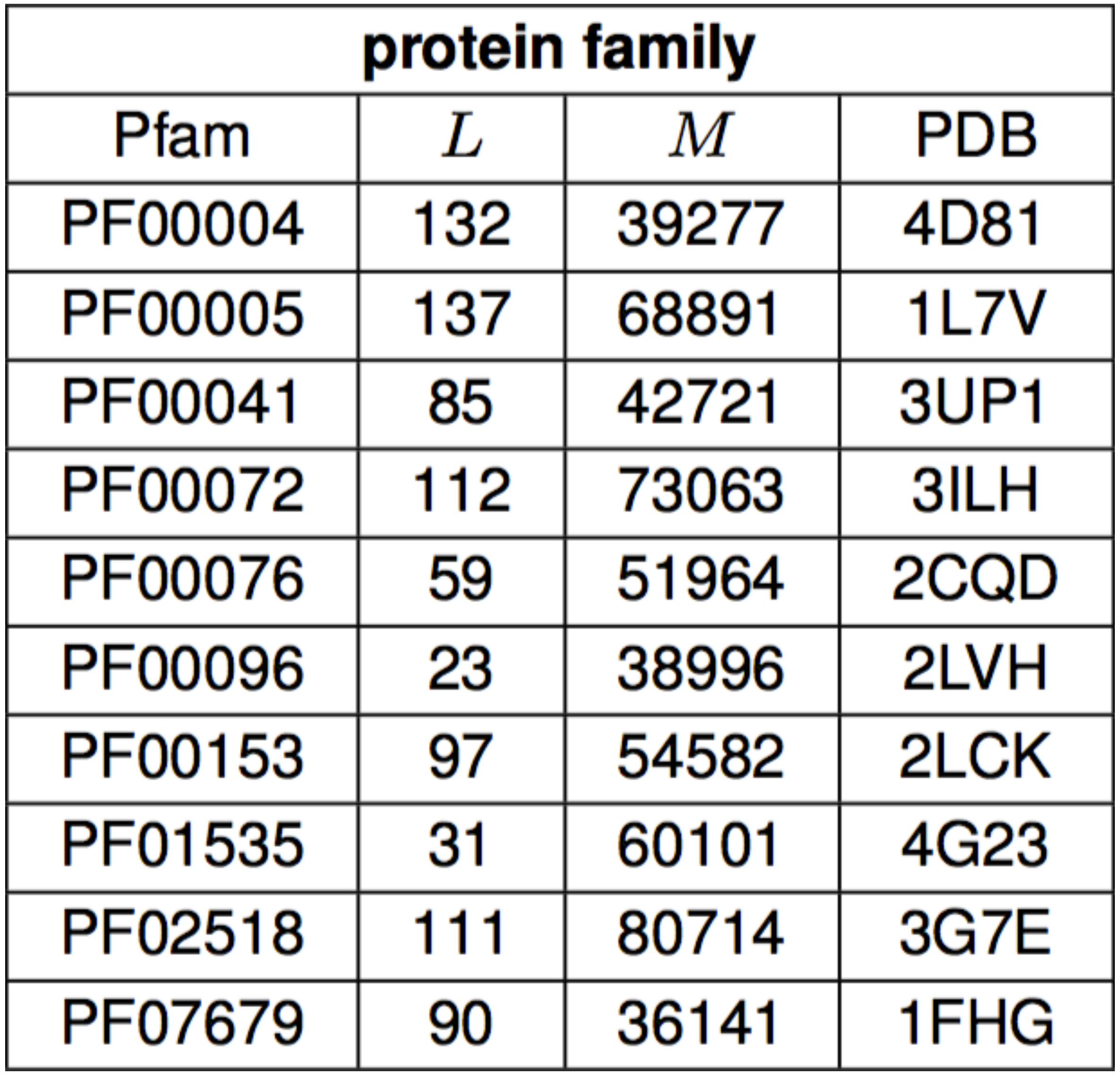
Black box modelization?

Understanding the model

Highly accurate implementation of the inference

Boltzmann Machine Learning (BM)

Learned on the **10 largest pfam families**



protein family			
Pfam	L	M	PDB
PF00004	132	39277	4D81
PF00005	137	68891	1L7V
PF00041	85	42721	3UP1
PF00072	112	73063	3ILH
PF00076	59	51964	2CQD
PF00096	23	38996	2LVH
PF00153	97	54582	2LCK
PF01535	31	60101	4G23
PF02518	111	80714	3G7E
PF07679	90	36141	1FHG

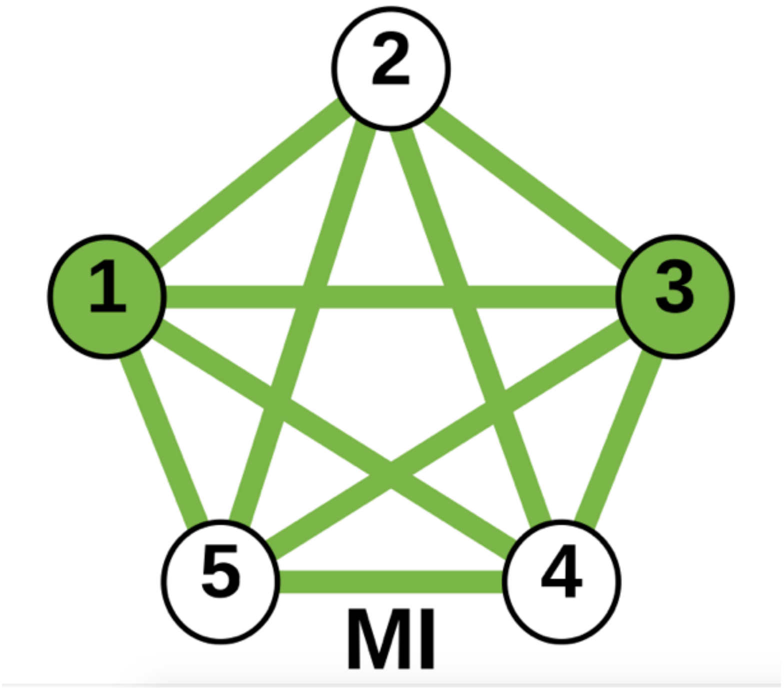
Analysis of the **indirect effects**

- Network of direct couplings?
- Biological interpretation?

Limitations of the model?

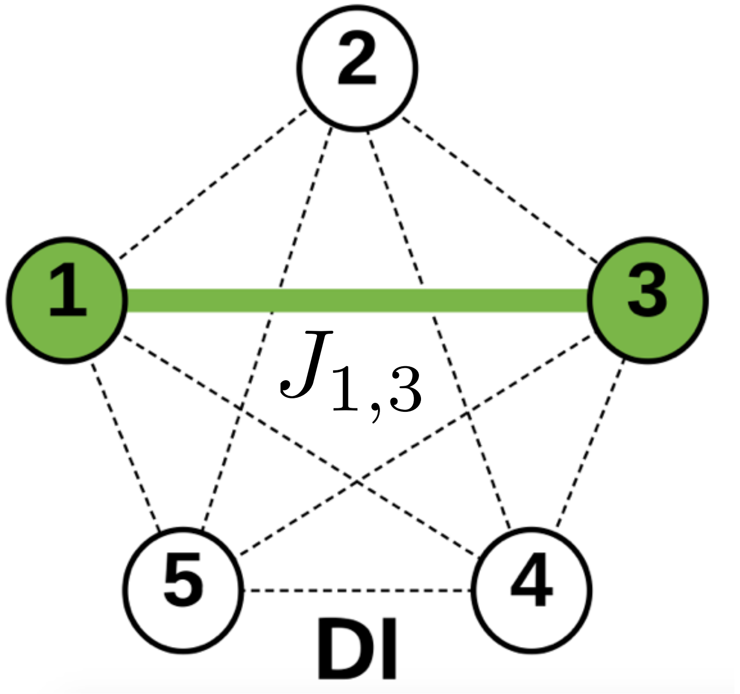
- Reproducing non-fitted features of the data?
- Need of higher order couplings?

Analysis of indirect effects



Mutual Information

Quantifying direct effects

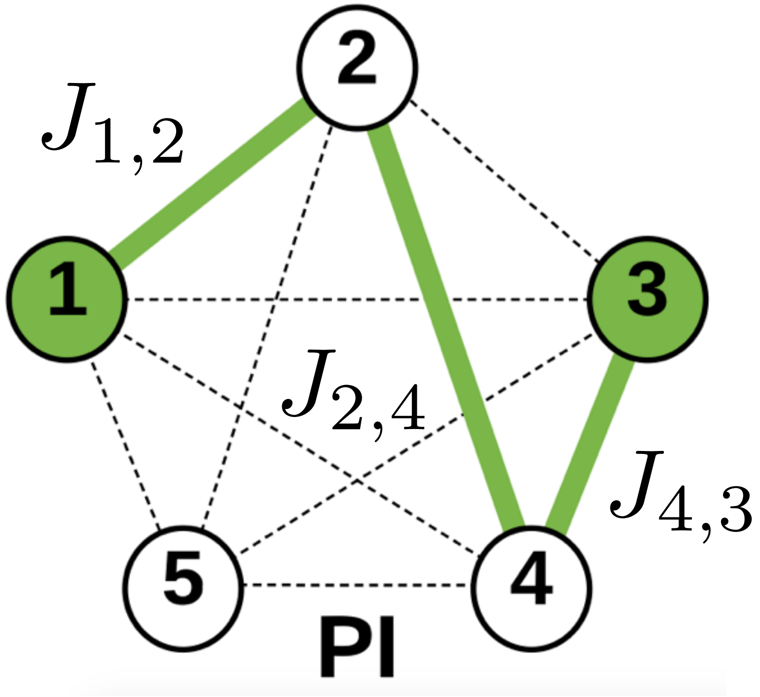


Direct Information

Strength of the direct coupling

Quantifying indirect effects

→ Chain of direct couplings!



Path Information

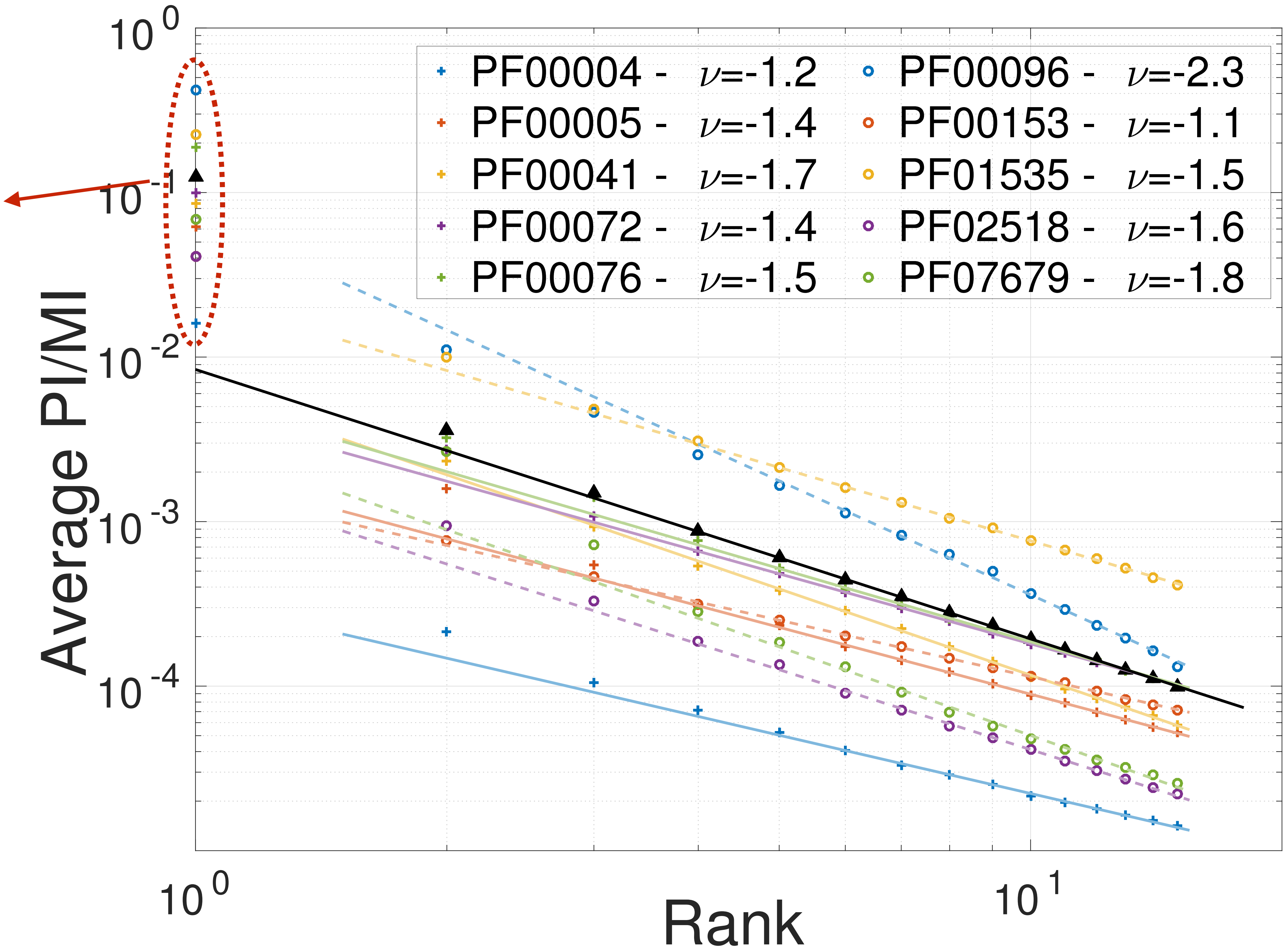
Effective coupling for a path

Collective effects?

How fast does path info. decrease?

$$\langle PI(rank)\rangle \propto rank^{-\nu}$$

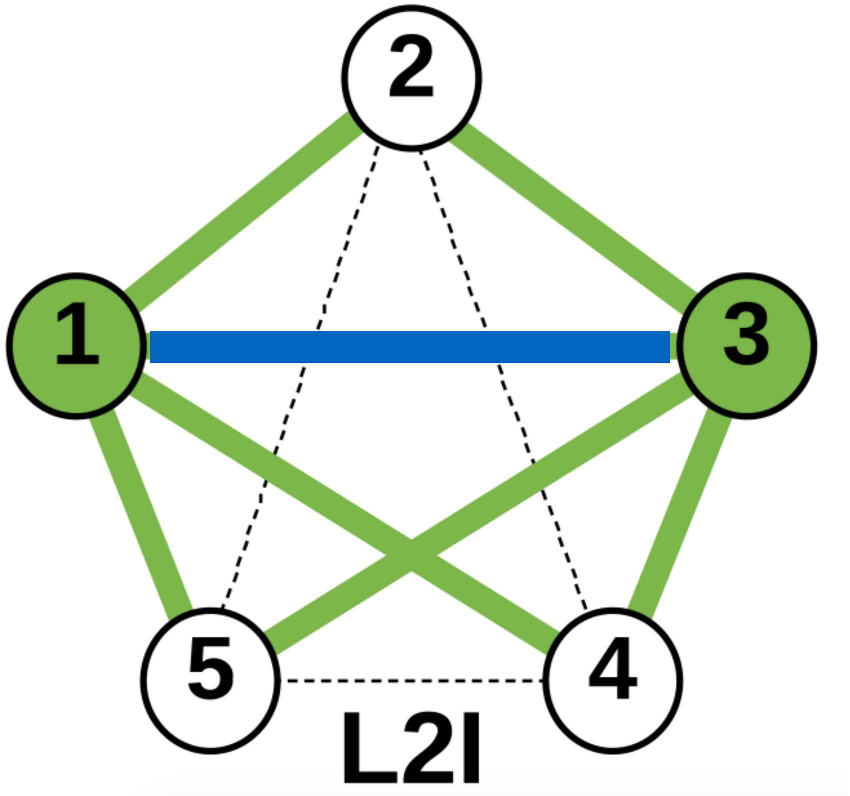
Computed on ~100 strongly correlated pairs



Strongest path
~
Direct path

Collective effect of numerous paths

We need to combine multiple paths!



Paths of length 2 are independent

$$P_2^{ij} \propto P_{ij}^{dir} \cdot \prod_{k \neq i,j} P^{path}\left([i\ k\ j]\right)$$

$\longrightarrow$ **Length 2 Information**

Effective coupling for all paths of length 2 (+ direct)

Strong direct coupling $\longrightarrow$ **contact**

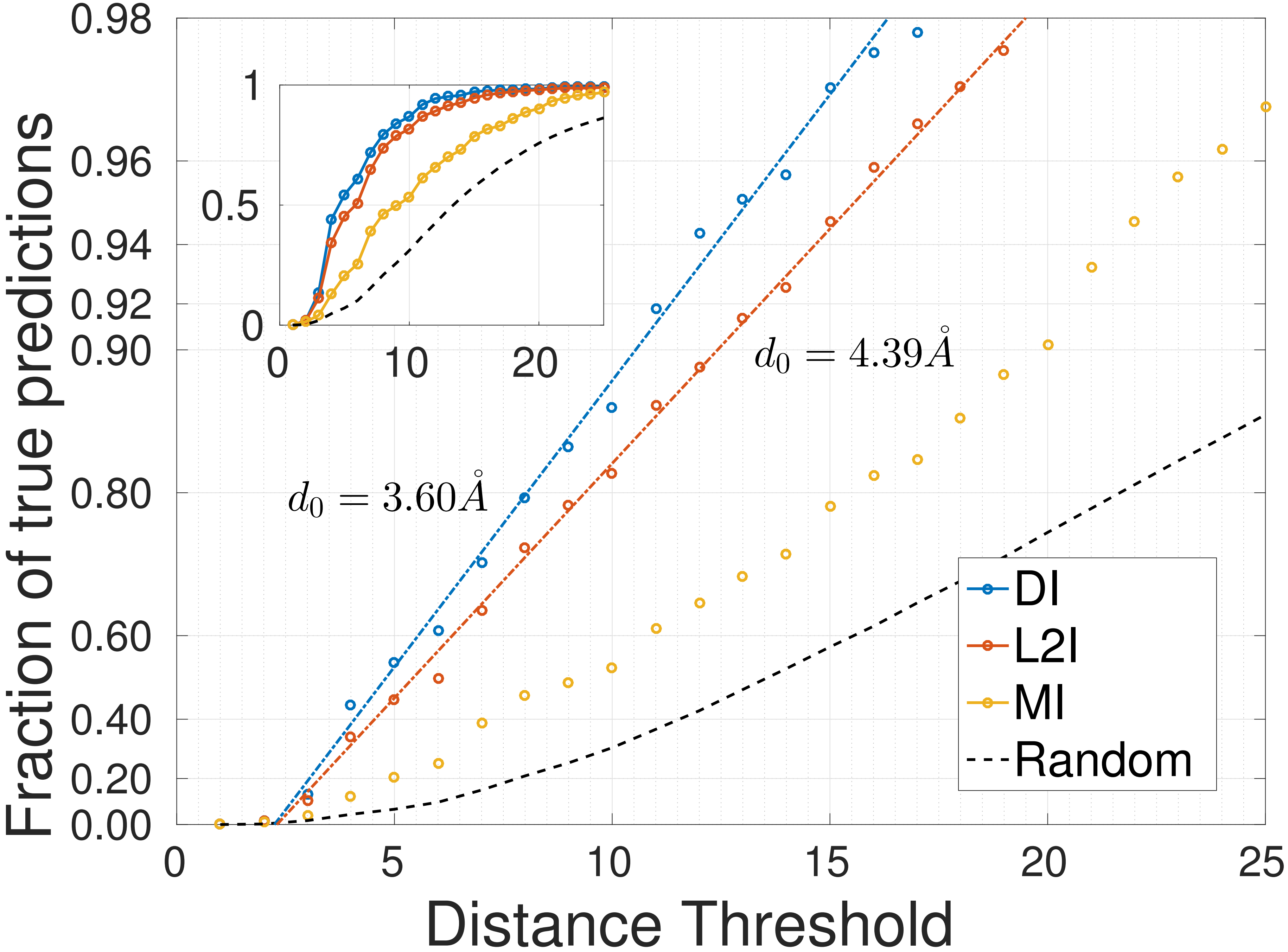
Strong 'length 2' effect $\xrightarrow{?}$ **Two contacts away**

Geometrical interpretation of indirect effects

Predicting proximity using different scores

Fit of the form $PPV = 1 - \exp\left(-d/d_0\right)$

Fraction of the 25 top scoring pairs distant of less than d angstroms



Compatible with **second structural neighbor**!

Limitations of the DCA model?

- How well does the DCA model capture information in the alignment?

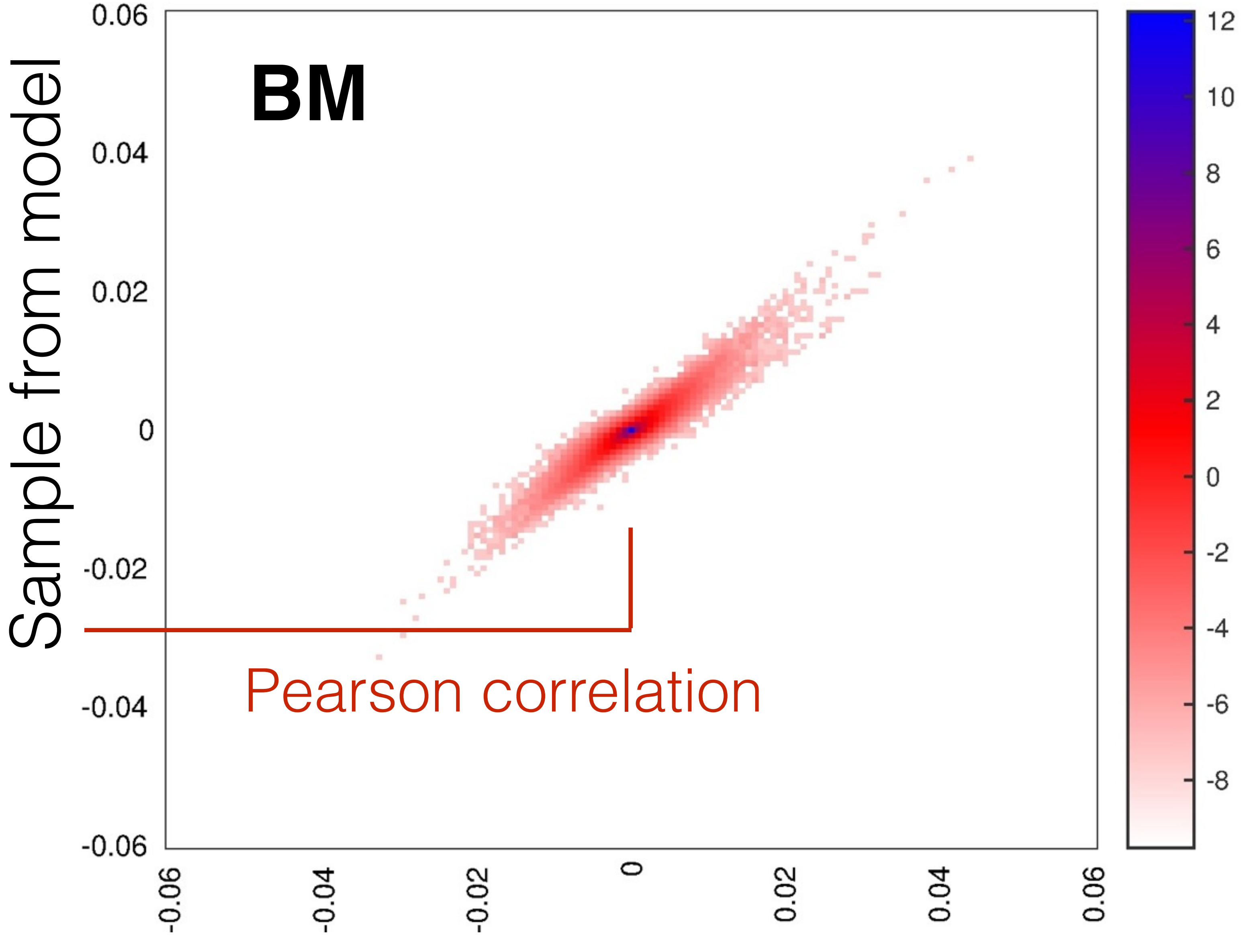
- Does one need higher order couplings to fully describe statistical features of the data?

→ Compare observables which are **not a direct consequence of the fitting procedure**!

Three points connected correlations

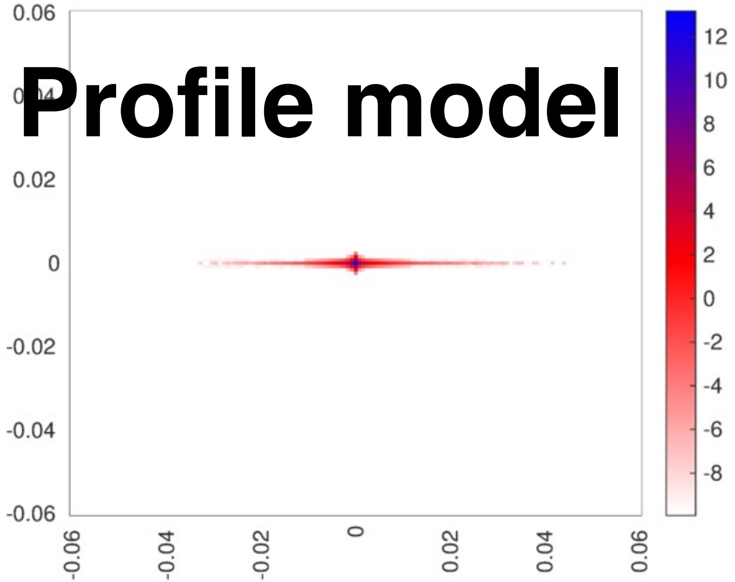
For PF00072

	three-point correlations	
Pfam	PLM	BM
PF00004	0.333	**0.980**
PF00005	0.718	**0.978**
PF00041	0.893	**0.991**
PF00072	0.803	**0.988**
PF00076	0.963	**0.993**
PF00096	ND	ND
PF00153	0.517	**0.986**
PF01535	0.120	**0.996**
PF02518	-0.228	**0.986**
PF07679	0.797	**0.993**

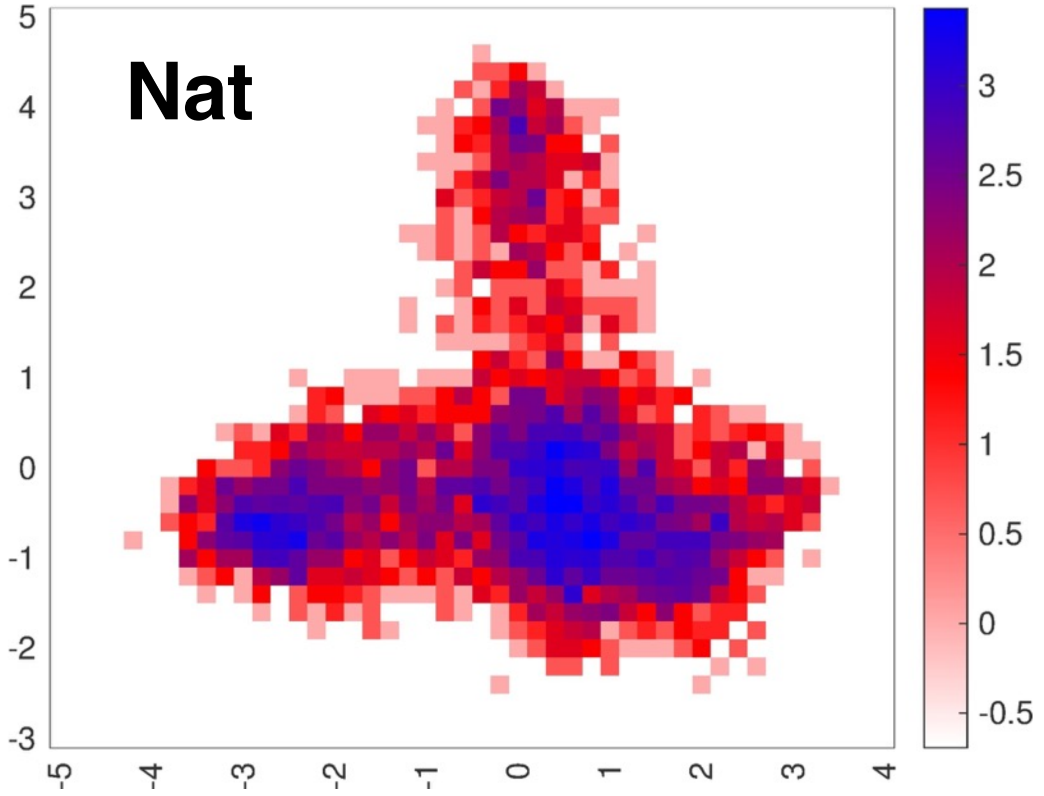


Pearson correlation

$$C_{ijk}(a,b,c) = f_{ijk}(a,b,c) - f_{ij}(a,b)f_k(c) \\ - f_{ik}(a,c)f_j(b) - f_{jk}(b,c)f_i(a) \\ + 2f_i(a)f_j(b)f_k(c)$$

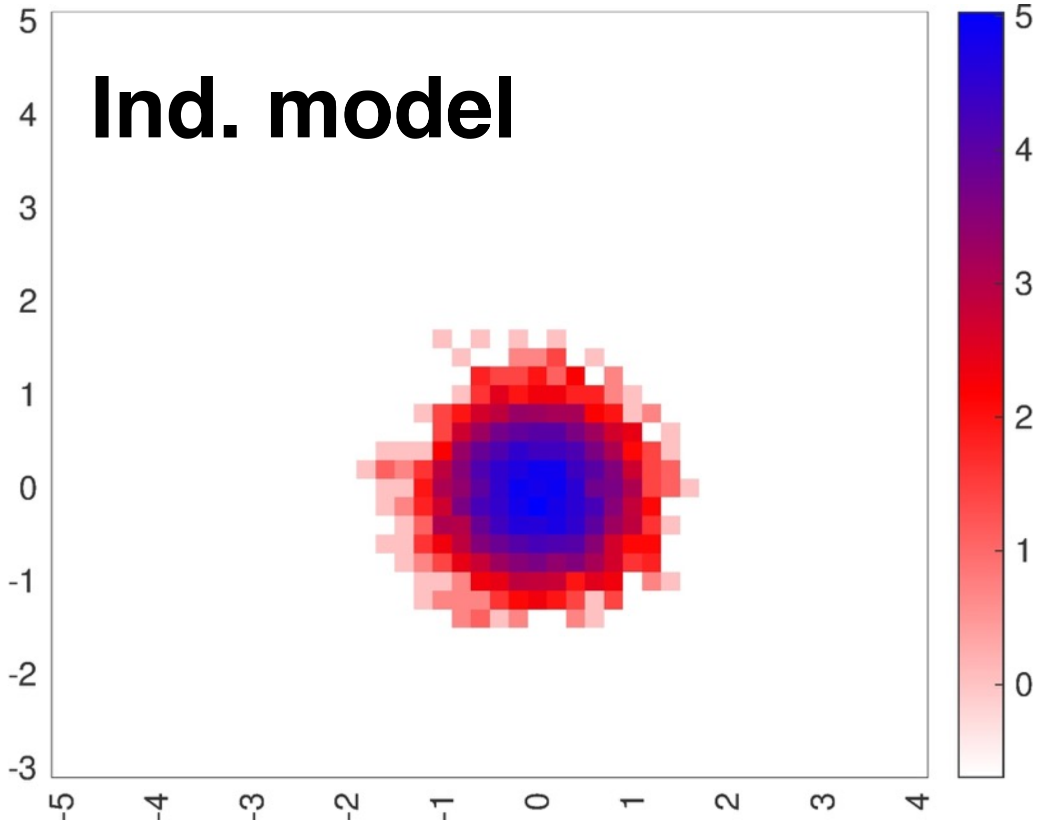


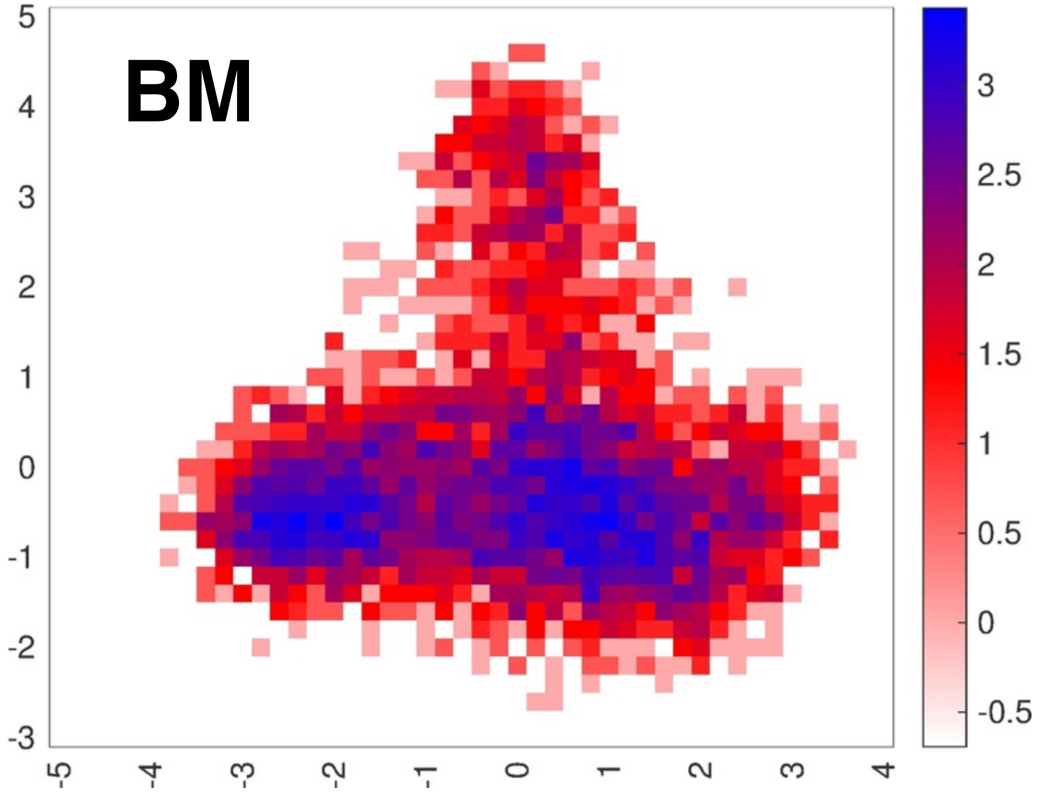
Sequences in principal component space



Projection of sequences on the first two principal components of the natural alignment

→ Higher order quantity





Limitations of the DCA model?

Inferred DCA models capture **non-fitted statistical features** of the natural sequences

- Three points connected correlations
- Global quantities (projection on PC’s, hamming distance distribution)

Pairwise couplings appear sufficient to capture variability in sequences of a protein family!

… which opens the way to protein design.

Limitations of the DCA model?

Inferred DCA models capture **non-fitted statistical features** of the natural sequences

- Three points connected correlations
- Global quantities (projection on PC's, hamming distance distribution)

Pairwise couplings appear sufficient to capture variability in sequences of a protein family!

… which opens the way to protein design.

Thank you!