Data-driven models of sequence landscapes

Martin Weigt

Laboratoire de Biologie Computationnelle et Quantitative
Sorbonne Université

Sanofi Campus Gentilly

4 Dec 2018

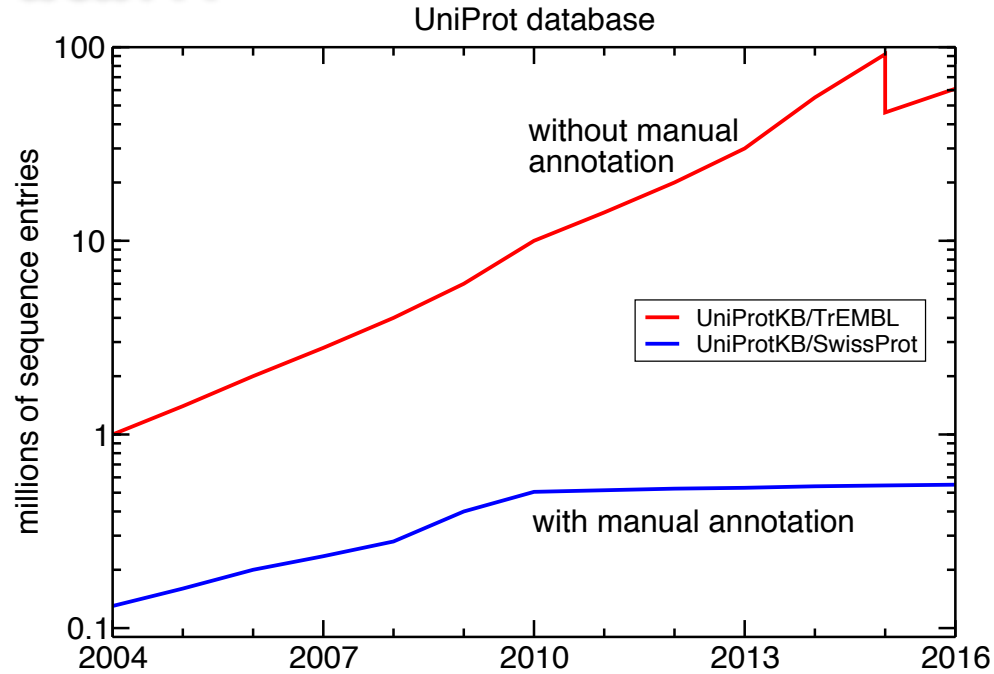
SCIENCES
SORBONNE
UNIVERSITÉ

All models are wrong, but some are useful.

[George E.P. Box, 1976]

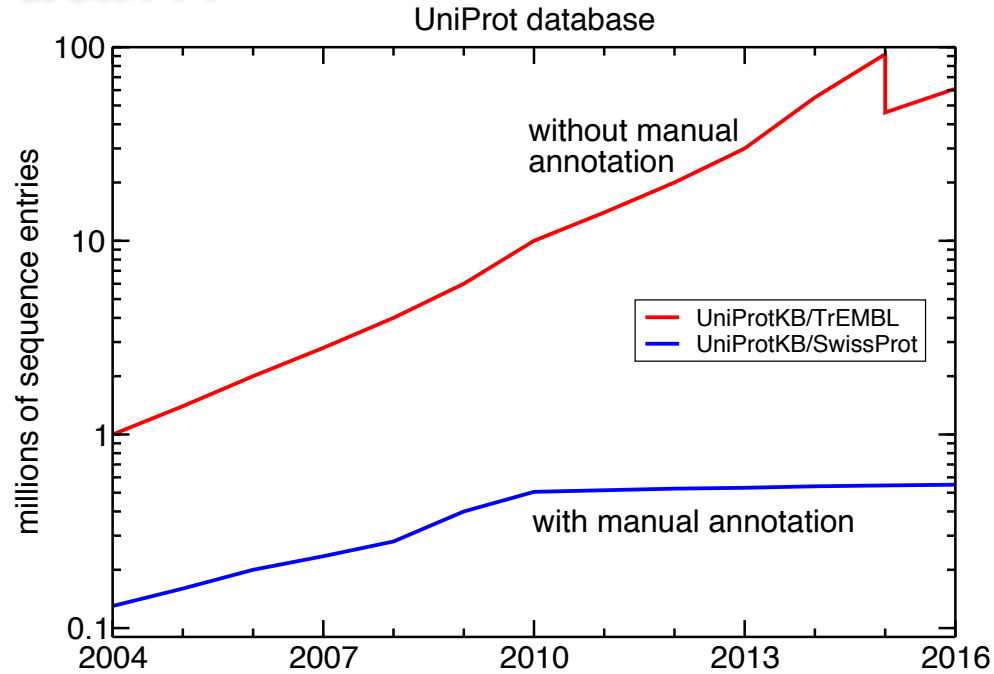
Data…

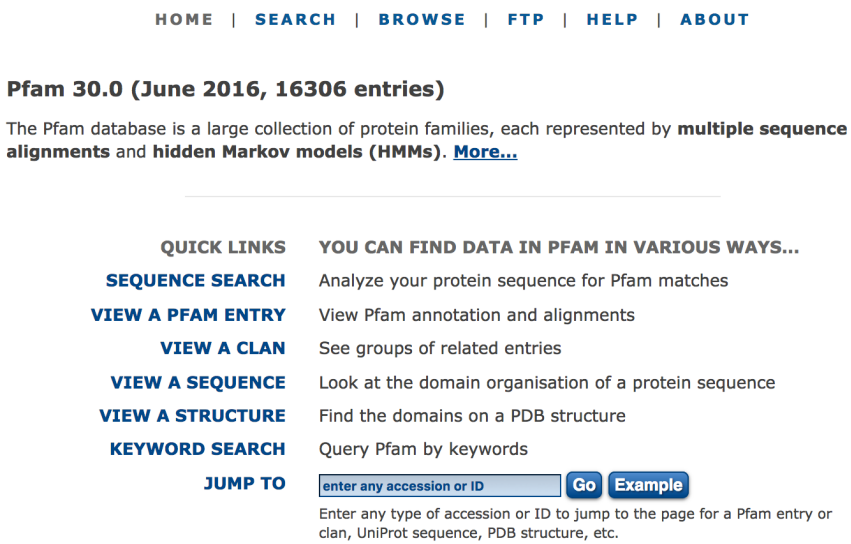
Sequence data are rapidly accumulating…



Data…

Sequence data are rapidly accumulating…



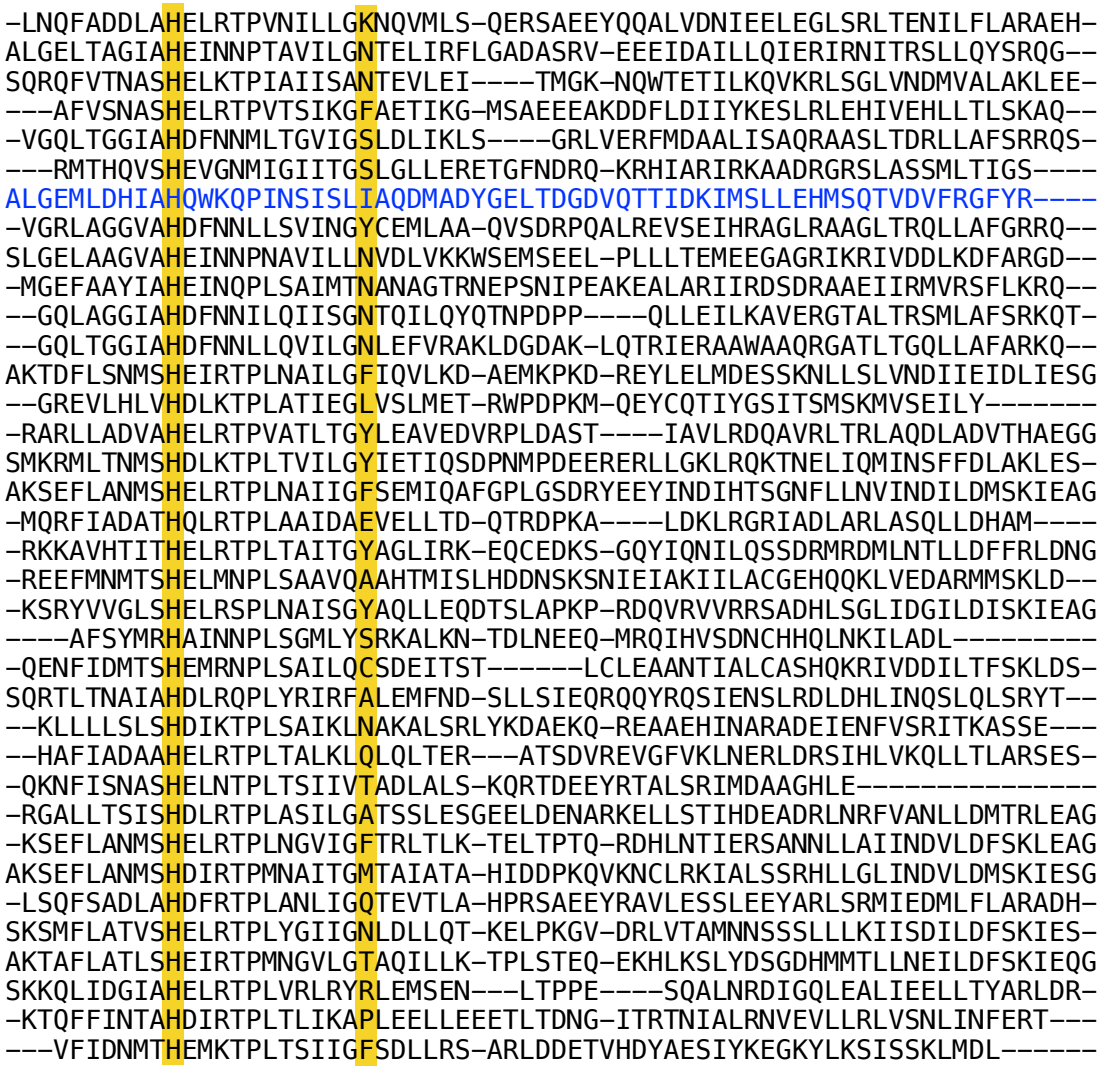


…and organized into **protein families**:

- common evolutionary origin (homologs)
- **conserved biological structure / function**
- **diverged sequences** (20-30% seq ID)
- available as multiple-sequence alignments

➡ **opportunity for data-driven approaches**

Looking for information in :



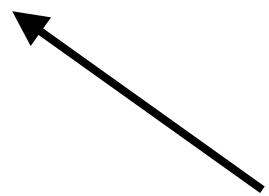
multiple-sequence alignment

- 50-500 positions sequence length
- 10^3-10^5 homologous sequences

Inference of statistical sequence models

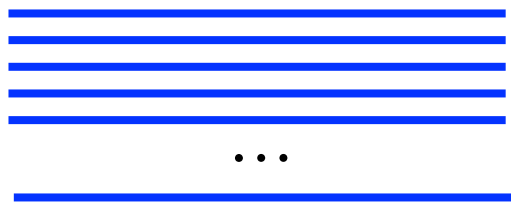
From data to statistical models

$$P(\overline{A}) = P(A_1, ..., A_L)$$



Data – MSA

$$\{\overline{A}^{\mu}\}_{\mu=1...M}$$



Attention:
Data insufficient to do this without modeling

Inference of statistical sequence models

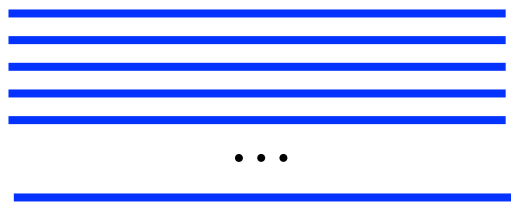
From data over observables to statistical models

$$P(\overline{A}) = \frac{1}{Z} \exp\left\{ \sum_a \lambda_a \mathcal{O}^a(\overline{A}) \right\}$$

maximum entropy model

Data – MSA

$$\{\overline{A}^\mu\}_{\mu=1\ldots M}$$



$$\langle \mathcal{O}^a(\overline{A}) \rangle_P = \frac{1}{M} \sum_\mu \mathcal{O}^a(\overline{A}^\mu)$$

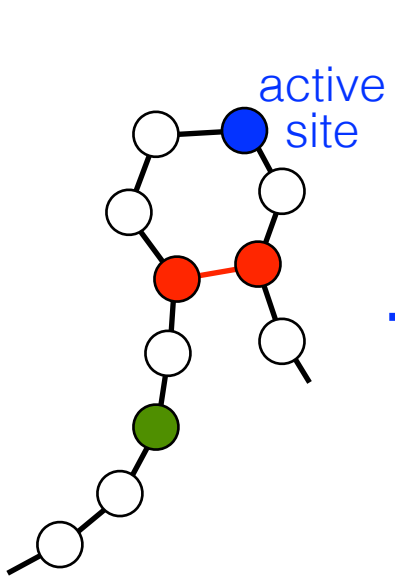
model average

average over data

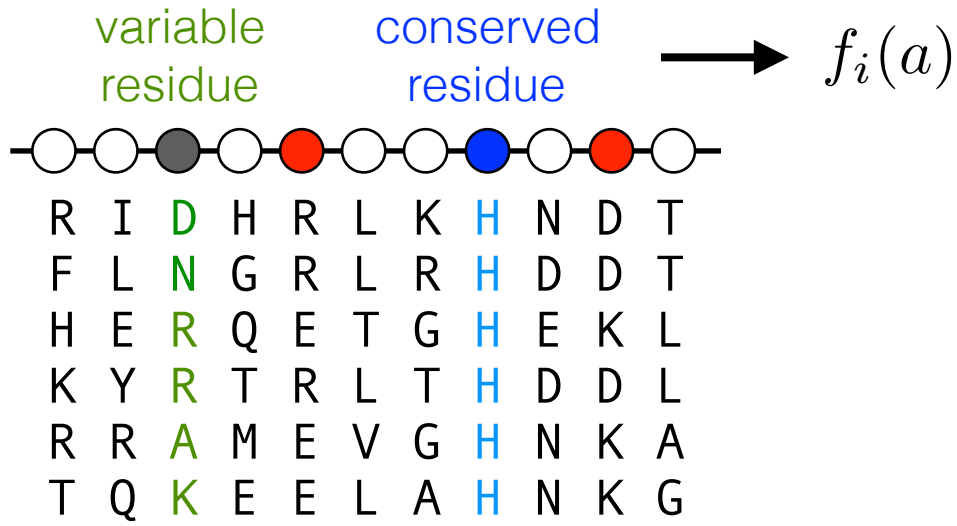
Attention:

- model depends on data only via sample-averaged observables
- selection of observables requires prior biological knowledge

Conservation in proteins



evolution



Profile model – independent residues

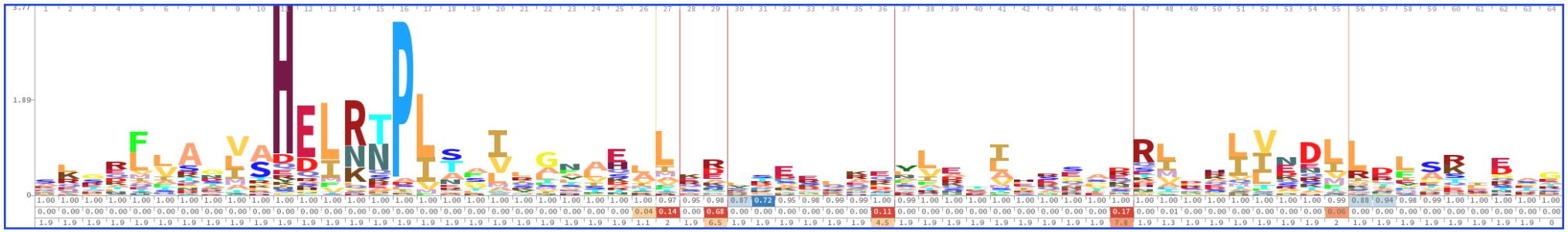
$$P(a_1, ..., a_L) \sim \exp\left\{\sum_i h_i(a_i)\right\}$$

statistical modeling

All models are wrong, but some are useful.

[George E.P. Box, 1976]

Profile models



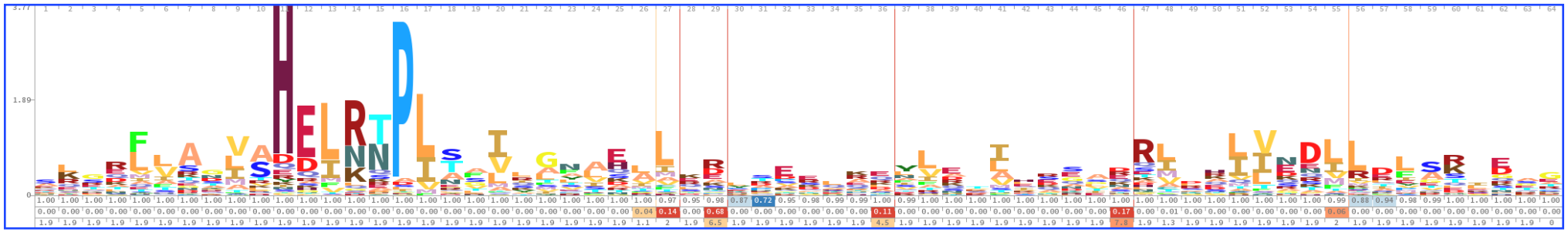
- identify conserved positions as functionally or structurally important positions
- are used to identify homologous sequences (automatic annotation)
- are used to align sequences

belong to the most successful statistical models in bioinformatics

All models are wrong, but some are useful.

[George E.P. Box, 1976]

Profile models



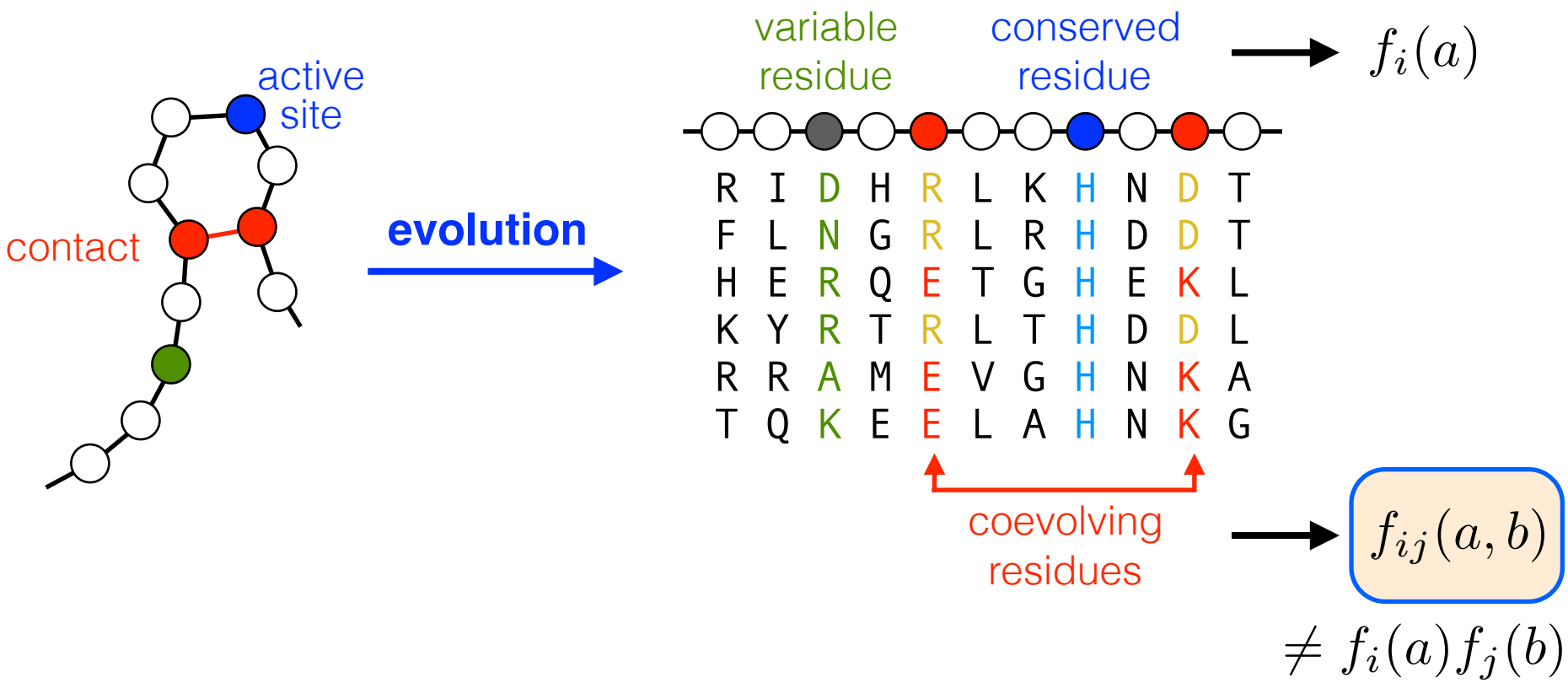
- identify conserved positions as functionally or structurally important positions
- are used to identify homologous sequences (automatic annotation)
- are used to align sequences

belong to the most successful statistical models in bioinformatics

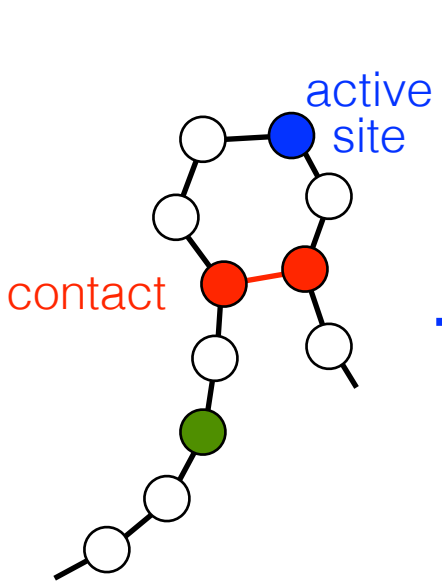
- are intrinsically unable to capture **relations** between positions
  - residue-residue contacts
  - protein-protein interaction

miss important information contained in sequence alignments

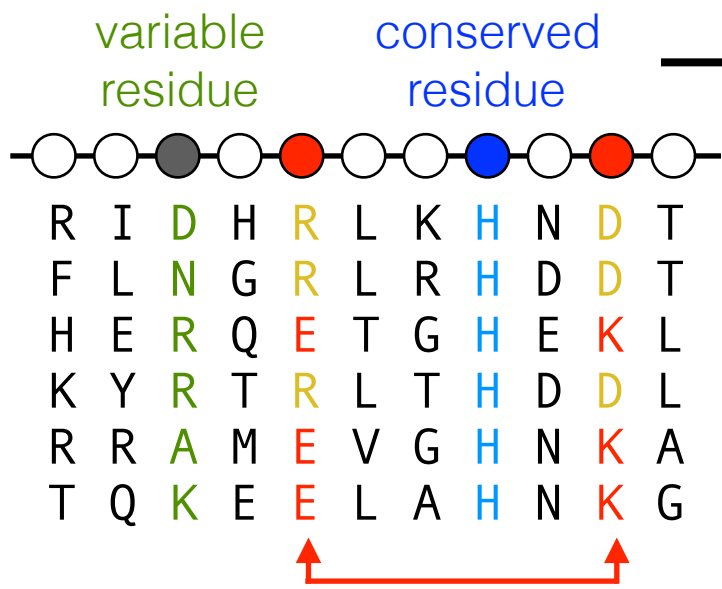
Conservation and coevolution in proteins



Conservation and coevolution in proteins



evolution



$f_i(a)$

$f_{ij}(a,b)$

$\neq f_i(a) f_j(b)$

Profile model – independent residues

$$P(a_1, ..., a_L) \sim \exp\left\{\sum_i h_i(a_i)\right\}$$

statistical modeling

Direct Coupling Analysis (DCA) – pairwise residue-residue couplings

$$P(a_1, ..., a_L)$$

$$\sim \exp\left\{\sum_{i<j} J_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right\}$$

[Weigt et al, PNAS '09]

[Morcos et al, PNAS '11]

Direct coupling analysis (DCA)

- Boltzmann-machine learning:
  - start with initialised parameters (fields/couplings)
  - calculate

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i,j\}} P(A_1, ..., A_L)$$

  - update parameters to fit marginals

$$\Delta J_{ij}(a, b) = \varepsilon [f_{ij}(a, b) - P_{ij}(a, b)]$$

  - iterate until sufficiently precise fitting

➡ exact calculation requires exponential time ~ 21^L

➡ approximations (MCMC) needed for computational efficiency

Direct coupling analysis (DCA)

- pseudo-likelihood maximisation
  - replace ensemble average by sample average over sequence data

$$P_1(A_1) = \sum_{\{A_i|i>1\}} P(A_1, ..., A_L)$$

$$= \sum_{\{A_i|i>1\}} P(A_1|A_2, ..., A_L) P(A_2, ..., A_L)$$

$$\simeq \frac{1}{M} \sum_{m=1}^{M} P(A_1|A_2^m, ..., A_L^m)$$

biological sequence data

[Balakrishnan et al., Proteins '11]
[Ekeberg, Lövkvist, Lan, MW, Aurell PRE '13]

Are pairwise DCA couplings useful?

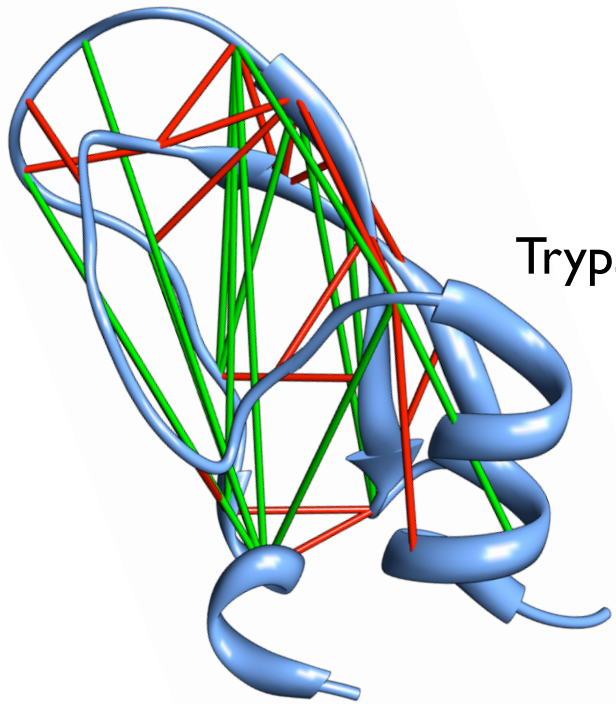
DCA models are **graphical statistical models**:

- defined on a **network** of strongly coupled residues
- provide a **probability** to each sequence (sequence landscape)

What is the **biological information** contained in such models ?

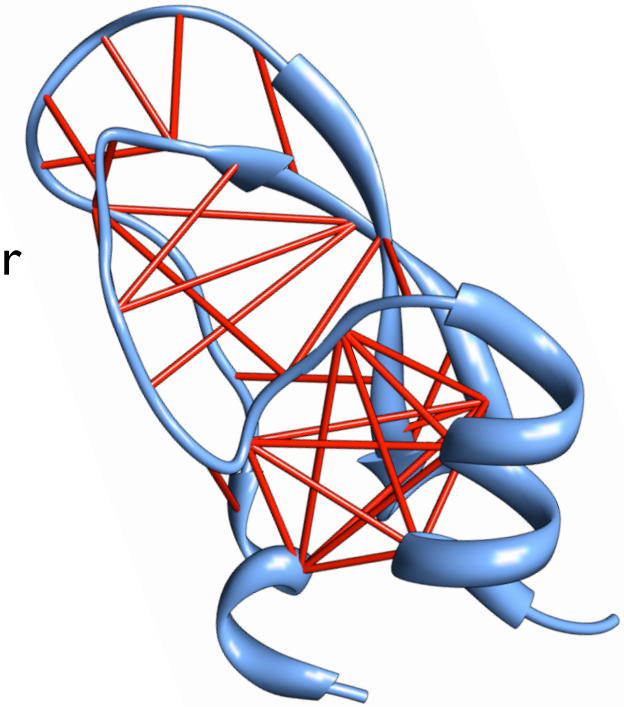
Strong couplings predict residue contacts

30 strongest correlations



Trypsin inhibitor

30 strongest couplings

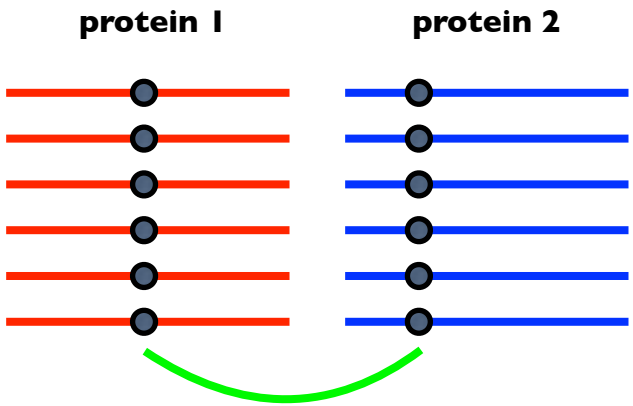


- works across numerous protein families
- accurate prediction for >1000 diverged sequences
- guides 3D protein structure prediction

[Marks et al., PLoS ONE '11]
[Sulkowska, Morcos, MW, Hwa, Onuchic, PNAS '12]
[Hopf et al., Cell '12]
[Nugent, Jones, PNAS '12]
[Ovchinnikov et al., eLife '15]
[Ovchinnikov et al., Science '17]

Prediction of inter-protein residue coevolution

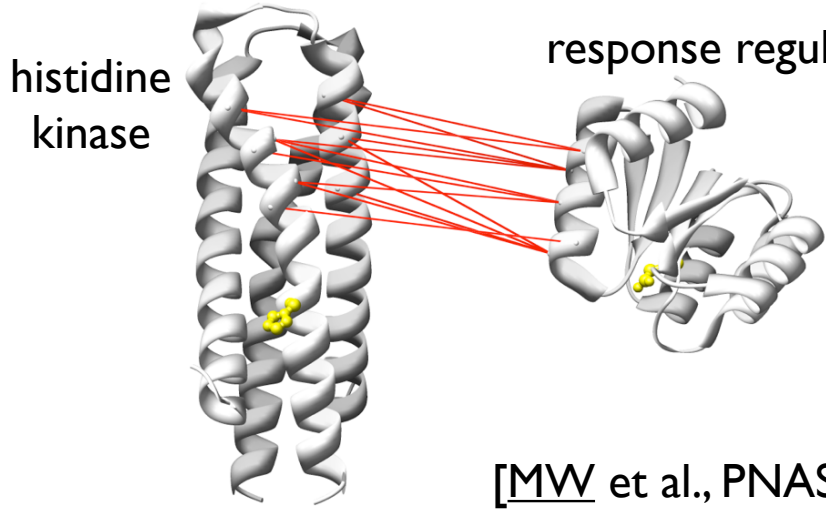
Joint MSA of protein families



inter-protein couplings

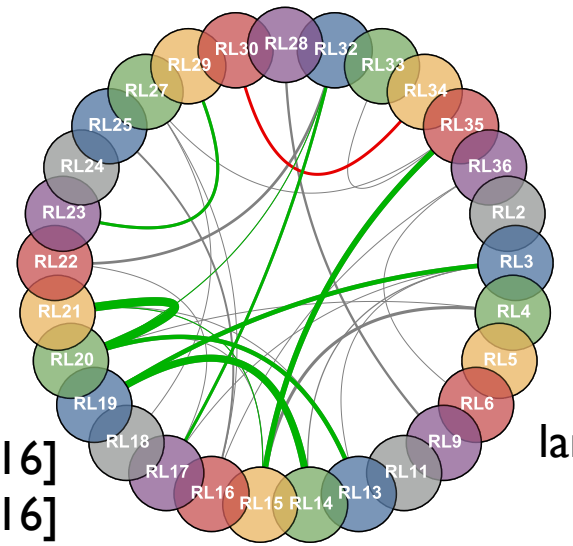
DCA

inter-protein residue contacts



[MW et al., PNAS '09]

protein-protein interaction networks



large ribosomal subunit

[Feinauer, Szurmant, MW, Pagnani, PLoS ONE '16]
[Gueudré, Baldassi, Zamparo, MW, Pagnani, PNAS '16]

Are pairwise DCA couplings useful?

DCA models are **graphical statistical models**:

- defined on a **network** of strongly coupled residues
- provide a **probability** to each sequence (sequence landscape)

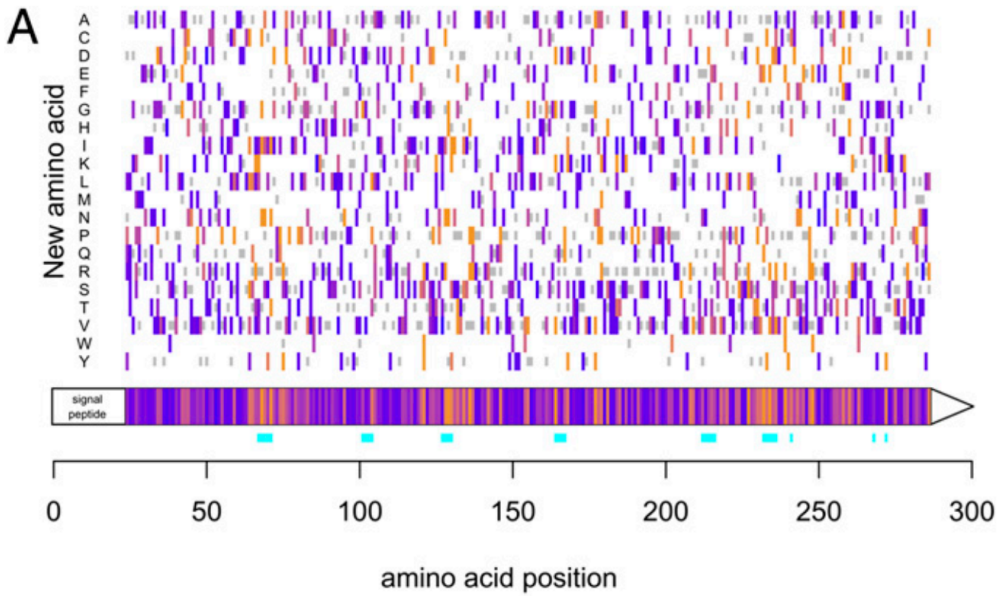
What is the **biological information** contained in such models ?

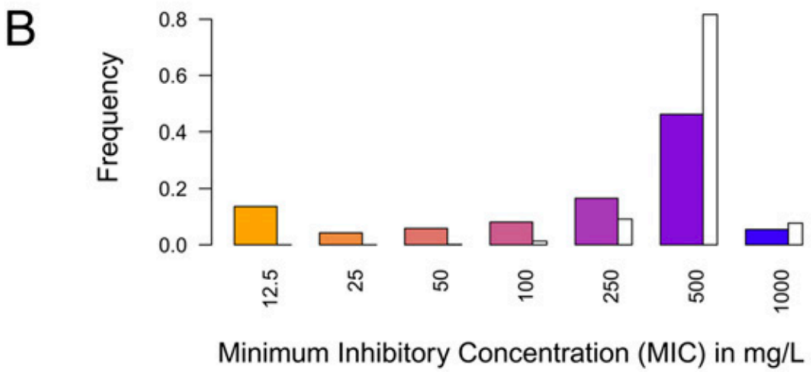
Measuring mutational effects in proteins

Capturing the mutational landscape of the beta-lactamase TEM-1

PNAS 110 (2013) 13067

Hervé Jacquier[a,b,c,1], André Birgy[a,b], Hervé Le Nagard[a,b,d,e], Yves Mechulam[f], Emmanuelle Schmitt[f], Jérémy Glodt[a,b], Beatrice Bercot[c,g], Emmanuelle Petit[h], Julie Poulain[h], Guilène Barnaud[i], Pierre-Alexis Gros[a,b,j], and Olivier Tenaillon[a,b,1]





Deep mutational scanning of proteins

TEM-1 protein

- causes antibiotic resistance

generated ~10^4 random mutants

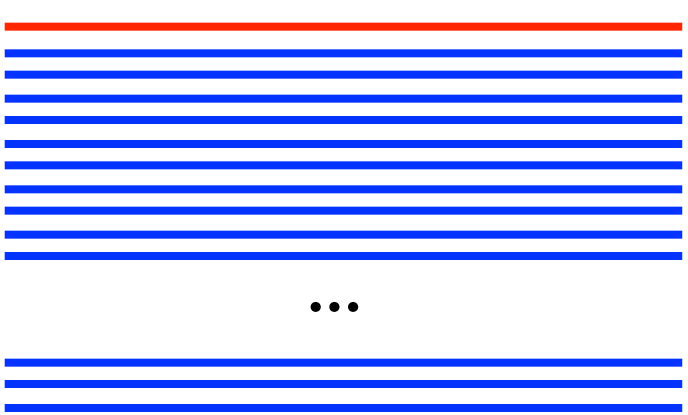
- 1,700 without mutation
- 990 distinct single AA changes

measured resistance to amoxicillin

- minimum inhibitory concentration as proxy for fitness

Landscape inference by Direct-Coupling Analysis

Beta-lactamase2 family (PF13354)



~2,500 diverged sequences

Statistical landscape inference (DCA)

$$P(A_1, ..., A_L)$$

$$\sim \exp\left\{ \sum_{i,j=1}^{L} e_{ij}(A_i, A_j) + \sum_{i=1}^{L} h_i(A_i) \right\}$$

Score for mutant AA sequences

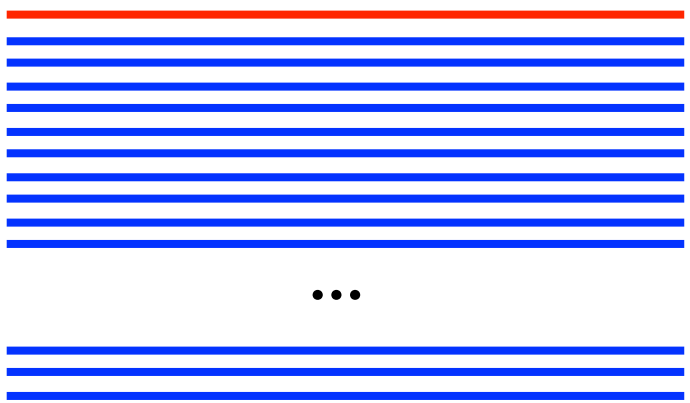
$$\Delta\Phi = \log\left\{ \frac{P(\text{mutant})}{P(\text{wildtype})} \right\}$$

Evolutionary constraints
across diverged homologs

[Figliuzzi, Jacquier, Schug, Tenaillon, MW, *Mol Biol Evol* '16]

Landscape inference by Direct-Coupling Analysis

Beta-lactamase2 family (PF13354)

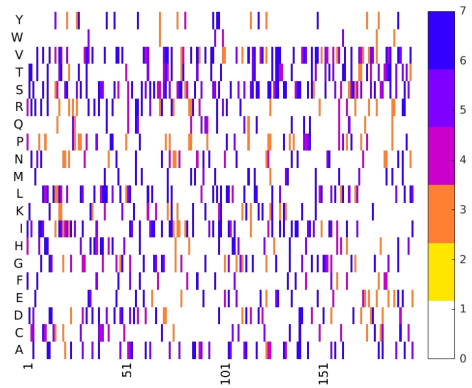


~2,500 diverged sequences

Statistical landscape inference (DCA)

$$P(A_1, ..., A_L)$$

$$\sim \exp\left\{ \sum_{i,j=1}^{L} e_{ij}(A_i, A_j) + \sum_{i=1}^{L} h_i(A_i) \right\}$$



MIC changes of TEM-1
due to single-AA changes

?

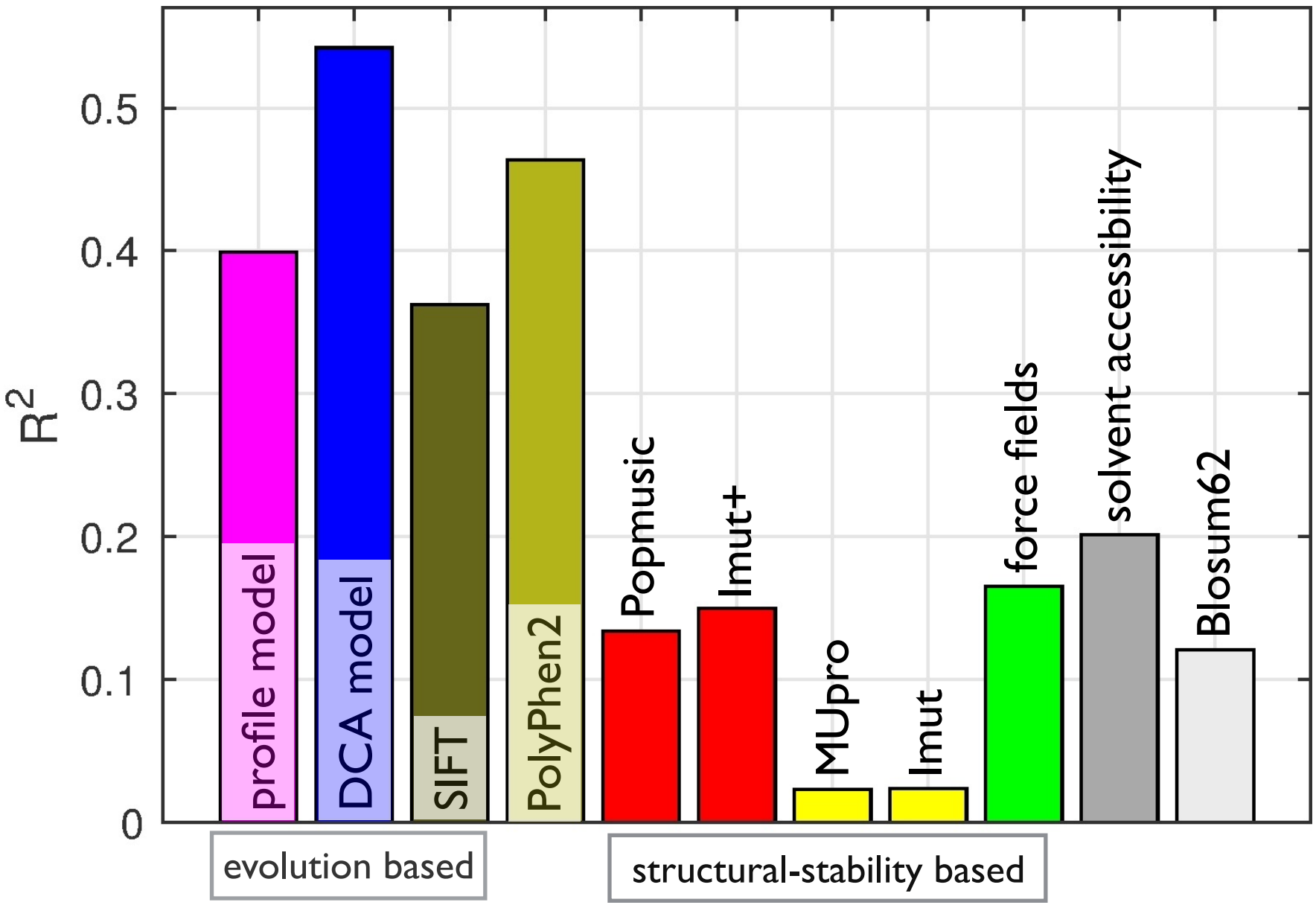
Score for mutant AA sequences

$$\Delta\Phi = \log\left\{ \frac{P(\text{mutant})}{P(\text{wildtype})} \right\}$$

Evolutionary constraints
across diverged homologs

[Figliuzzi, Jacquier, Schug, Tenaillon, MW, *Mol Biol Evol* '16]

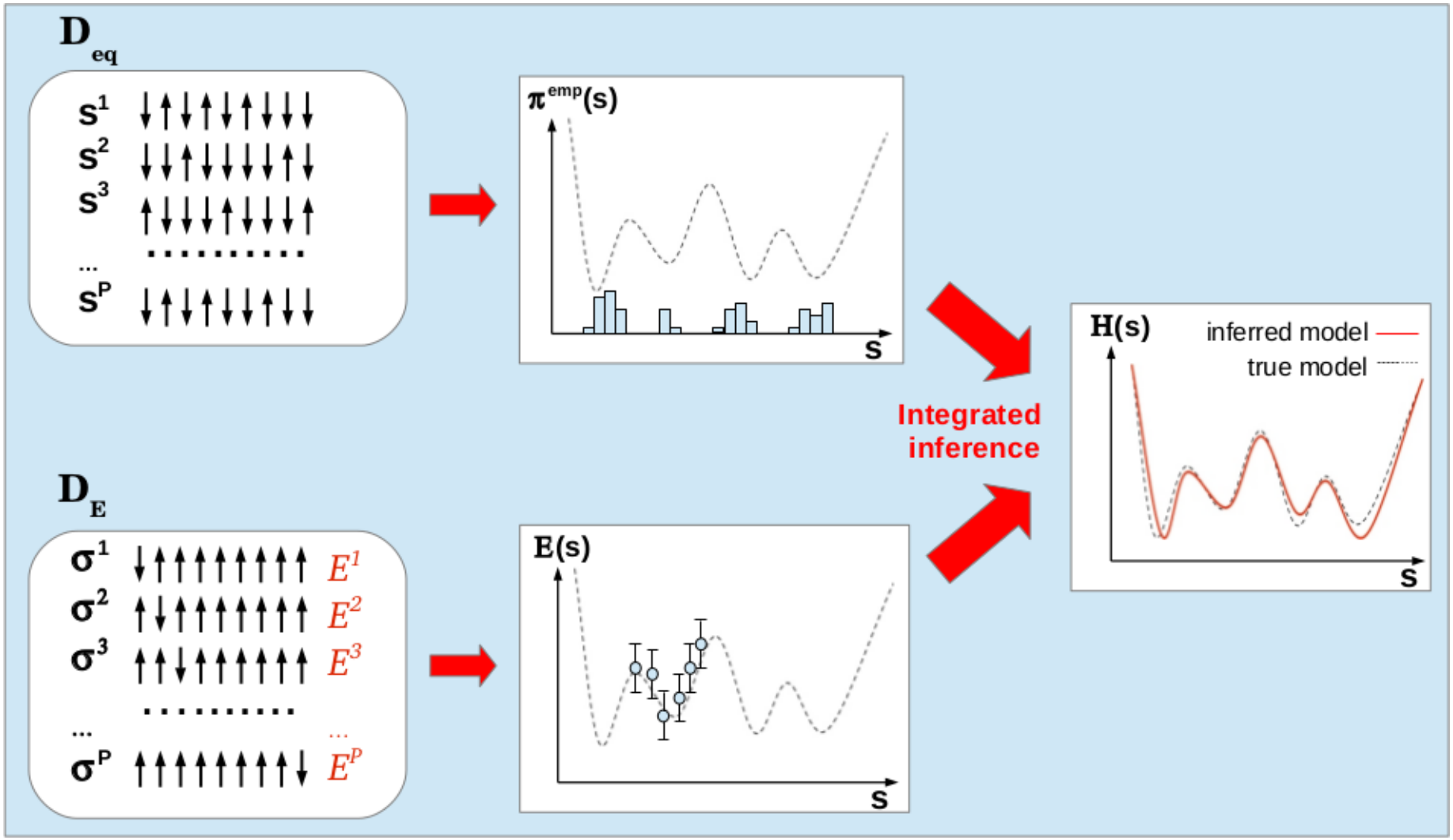
Predicting mutational effects in proteins



[Figliuzzi, Jacquier, Schug, Tenaillon, MW, *Mol Biol Evol* '16]

Integrating heterogenous data in landscape inference

MSA: sequence data



DMS: mutational data

[Barrat-Charlaix,Figliuzzi, MW, *Sci Rep* '16]

Integrating heterogenous data in landscape inference

Statistical model

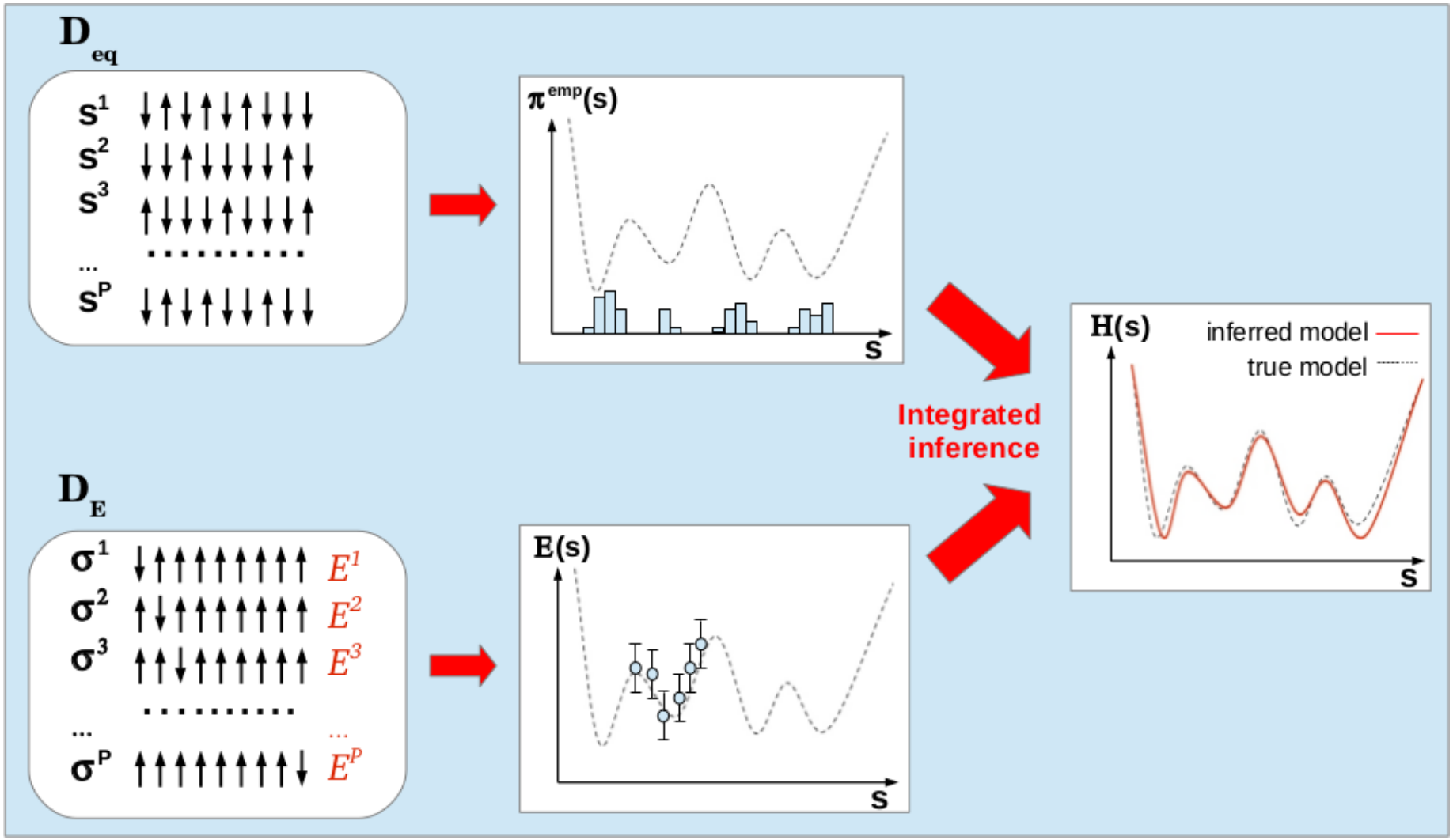
$$P(\overline{A}|\mathbf{J},\mathbf{h}) = \frac{1}{Z}\exp\left\{\sum_{a<b} J_{ij}(A_i,A_j) + \sum_a h_i(A_i)\right\} = \frac{1}{Z}\exp\left\{-\mathcal{H}(\overline{A})\right\}$$

Probability of MSA

$$P(\{\overline{A}^\mu\}_\mu|\mathbf{J},\mathbf{h}) = \exp\left\{-\sum_\mu \mathcal{H}(\overline{A}^\mu) - M\ln Z\right\}$$

Integrating heterogenous data in landscape inference

MSA: sequence data



DMS: mutational data

Integrating heterogenous data in landscape inference

Statistical model

$$P(\overline{A}|\mathbf{J},\mathbf{h}) = \frac{1}{Z}\exp\left\{\sum_{a<b} J_{ij}(A_i,A_j) + \sum_a h_i(A_i)\right\} = \frac{1}{Z}\exp\left\{-\mathcal{H}(\overline{A})\right\}$$

Probability of MSA

$$P(\{\overline{A}^\mu\}_\mu|\mathbf{J},\mathbf{h}) = \exp\left\{-\sum_\mu \mathcal{H}(\overline{A}^\mu) - M\ln Z\right\}$$

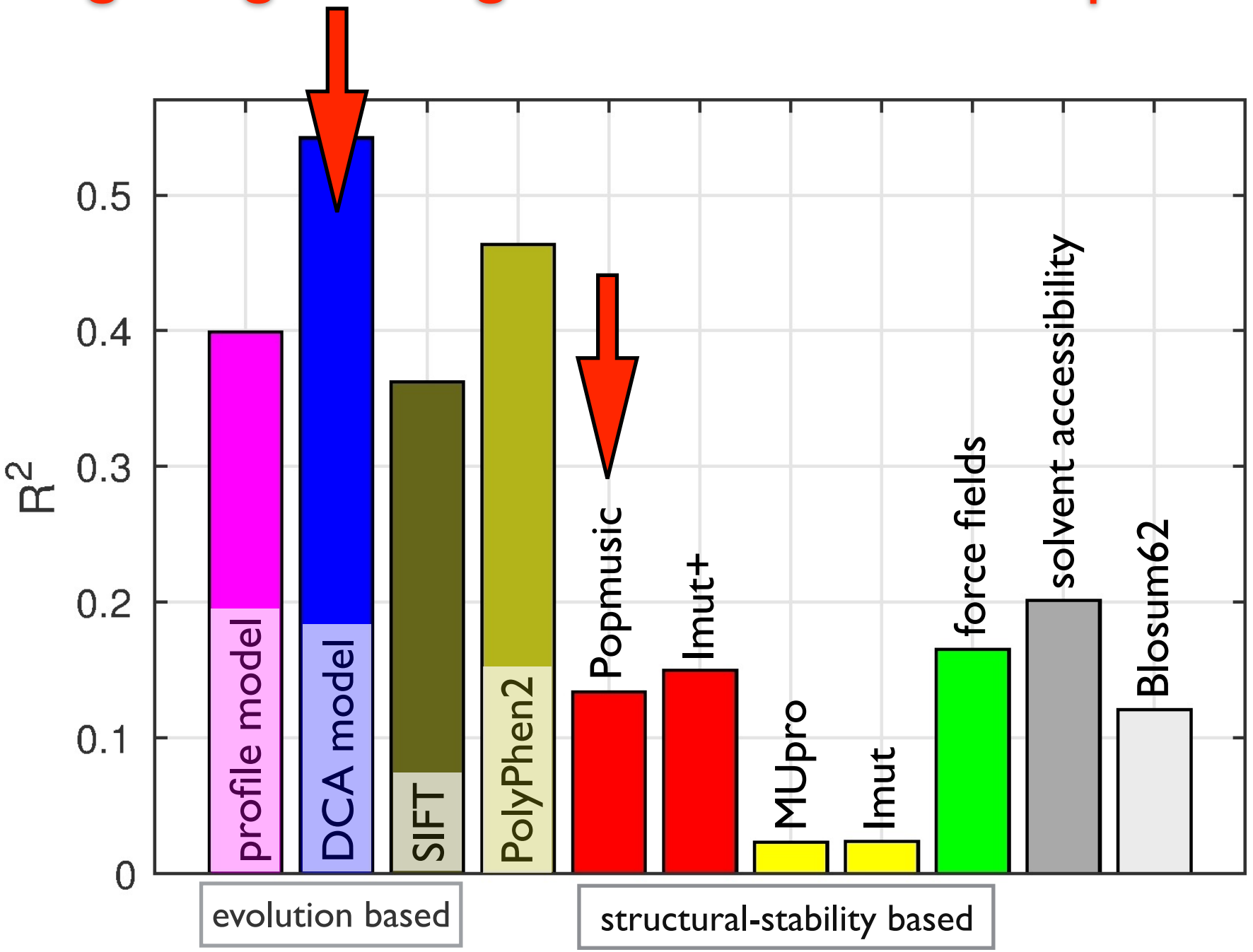
Mutational data - Gaussian experimental noise

$$P(\{E^a\}_a|\{\overline{A}^a\}_a,\mathbf{J},\mathbf{h})) = \frac{1}{(2\pi\Delta^2)^{P/2}}\exp\left\{-\frac{1}{2\Delta^2}\sum_a\left(E^a - \mathcal{H}(\overline{A}^a)\right)^2\right\}$$

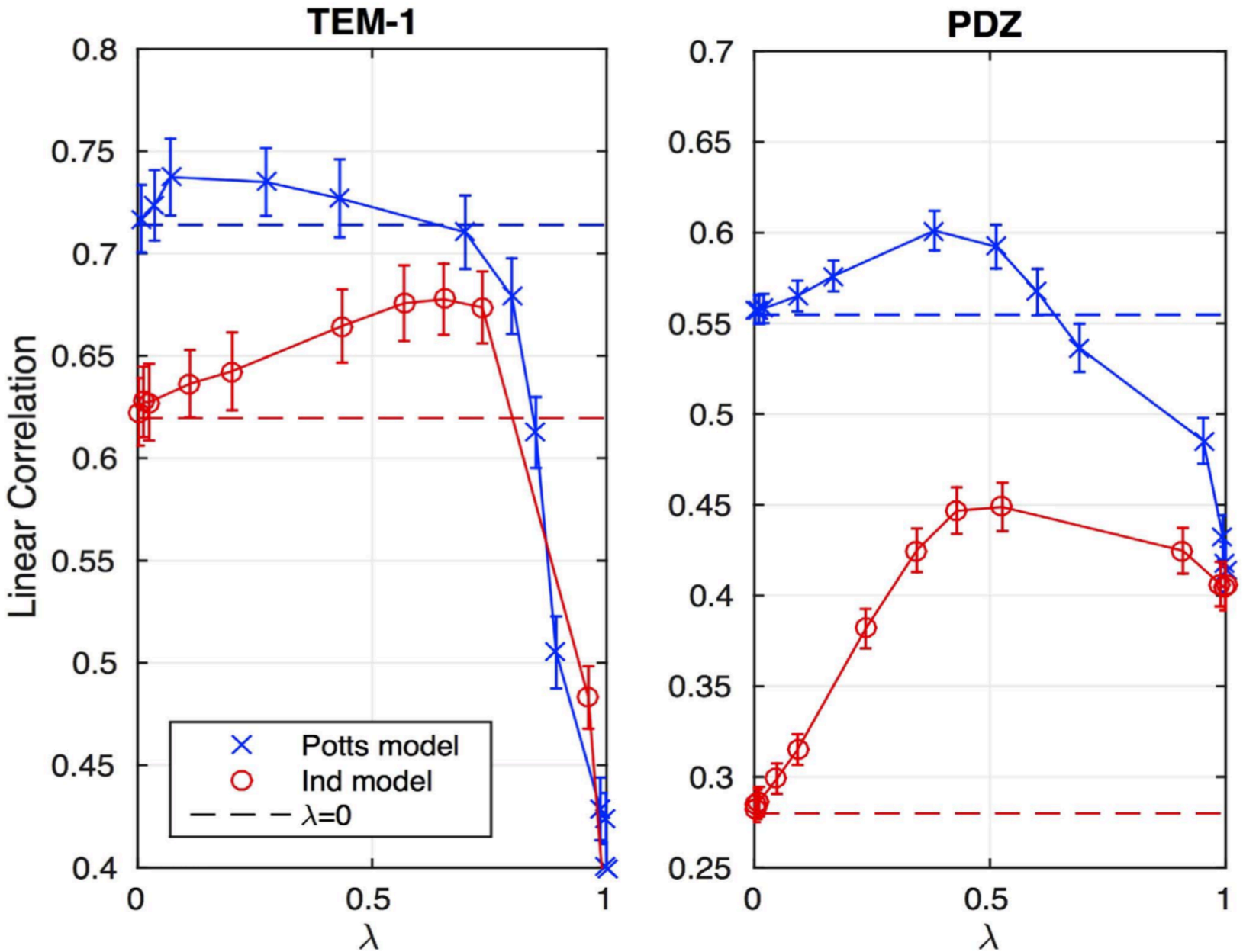
Log-likelihood of model parameters

$$\mathcal{L}(\mathbf{J},\mathbf{h}|data) = \log P(\{\overline{A}^\mu\}_\mu|\mathbf{J},\mathbf{h}) + \log P(\{E^a\}_a|\{\overline{A}^a\}_a,\mathbf{J},\mathbf{h})$$

Integrating heterogenous data in landscape inference



Integrating heterogenous data in landscape inference



[Barrat-Charlaix,Figliuzzi, MW, *Sci Rep* '16]

Are pairwise DCA couplings useful?

DCA models are **graphical statistical models**:

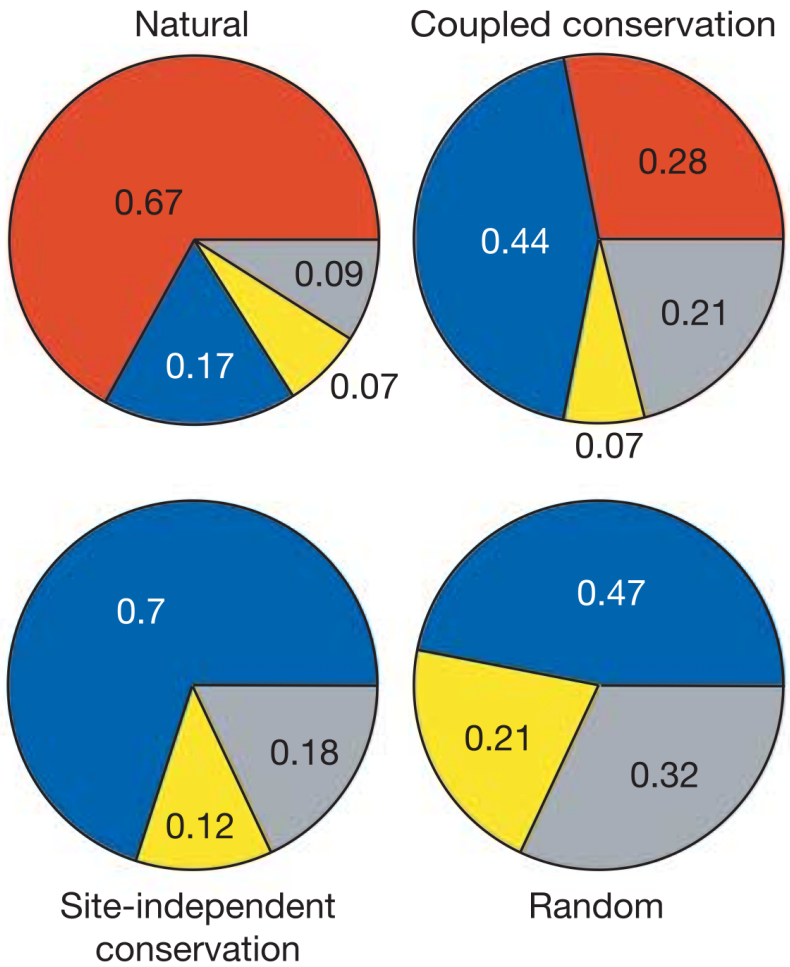
- defined on a **network** of strongly coupled residues
- provide a **probability** to each sequence (sequence landscape)

What is the **biological information** contained in such models ?

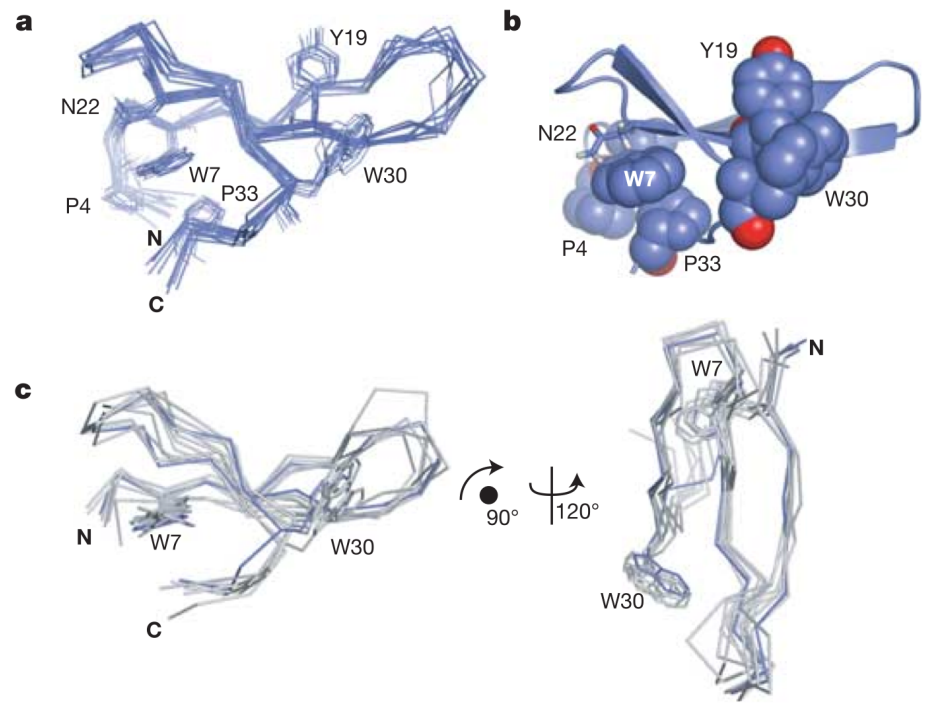
Are pairwise couplings sufficient?

Idea: scramble multiple-sequence alignments of homologs to conserve

- global amino-acid frequencies (no site specificity)
- site independent conservation (protein profile) → only local fields, no couplings
- pairwise amino-acid co-occurrences → pairwise residue-residue couplings



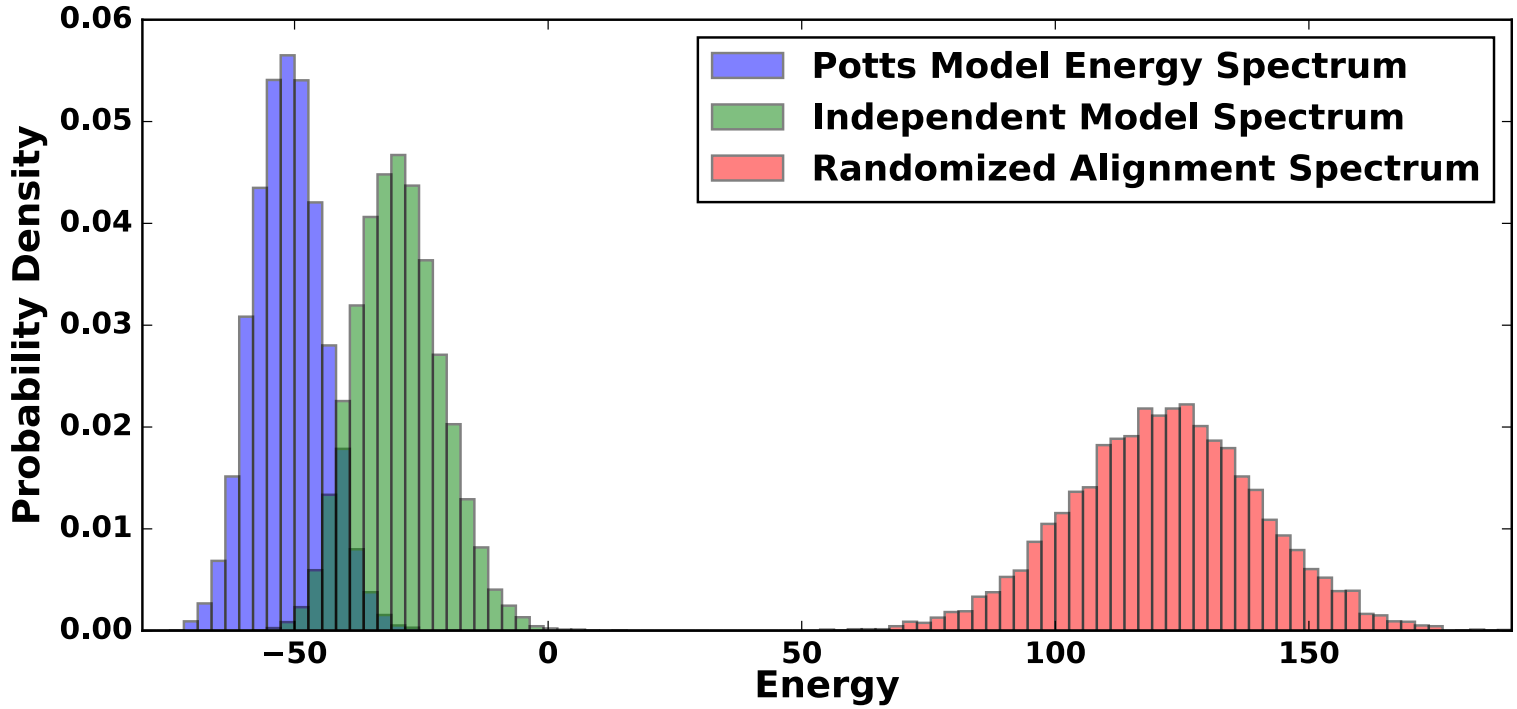
NMR structure of artificial protein (WW domain)

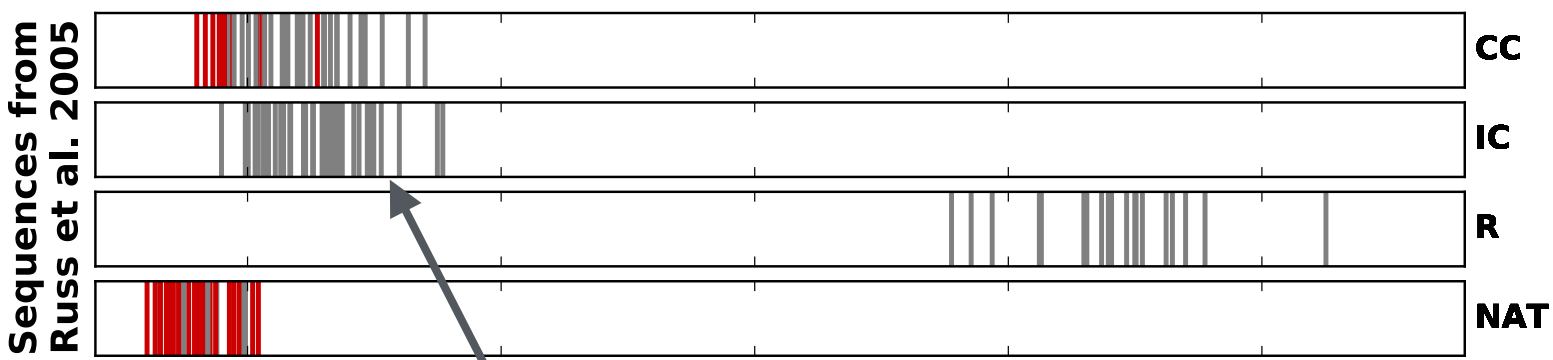


structural alignment with 6 natural WW domains

[M Socolich et al., *Nature* 437, 512 (2005)]

Predicting folding sequences of the WW domain





non folding

folding

high probability

low probability

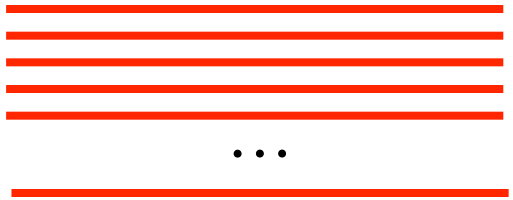
Data-driven statistical design

From data over observables and models to data

$$P(\overline{S}) \sim \exp\left\{\sum_a \lambda_a \mathcal{O}^a(\overline{S})\right\}$$

Output data:

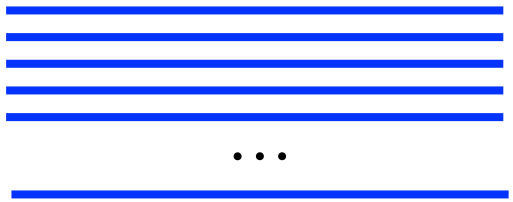
$$\{\overline{X}^\nu\}_{\nu=1,\ldots,N}$$



$$\langle \mathcal{O}_a(\overline{S})\rangle_P \simeq \frac{1}{M}\sum_\mu \mathcal{O}_a(\overline{S}^\mu)$$

Input data:

$$\{\overline{S}^\mu\}_{\mu=1,\ldots,M}$$



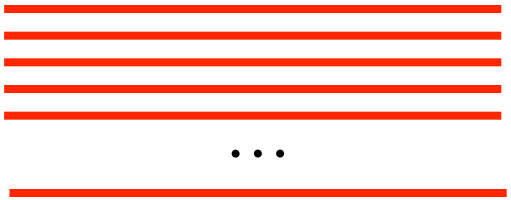
Data-driven statistical design

From data over observables and models to data

$$P(\overline{S}) \sim \exp\left\{\sum_a \lambda_a \mathcal{O}^a(\overline{S})\right\}$$

Output data:

$$\{\overline{X}^\nu\}_{\nu=1,\ldots,N}$$

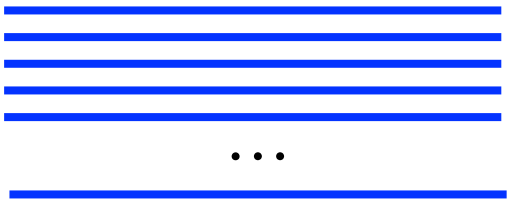


generative model?

$$\langle \mathcal{O}_a(\overline{S})\rangle_P \simeq \frac{1}{M}\sum_\mu \mathcal{O}_a(\overline{S}^\mu)$$

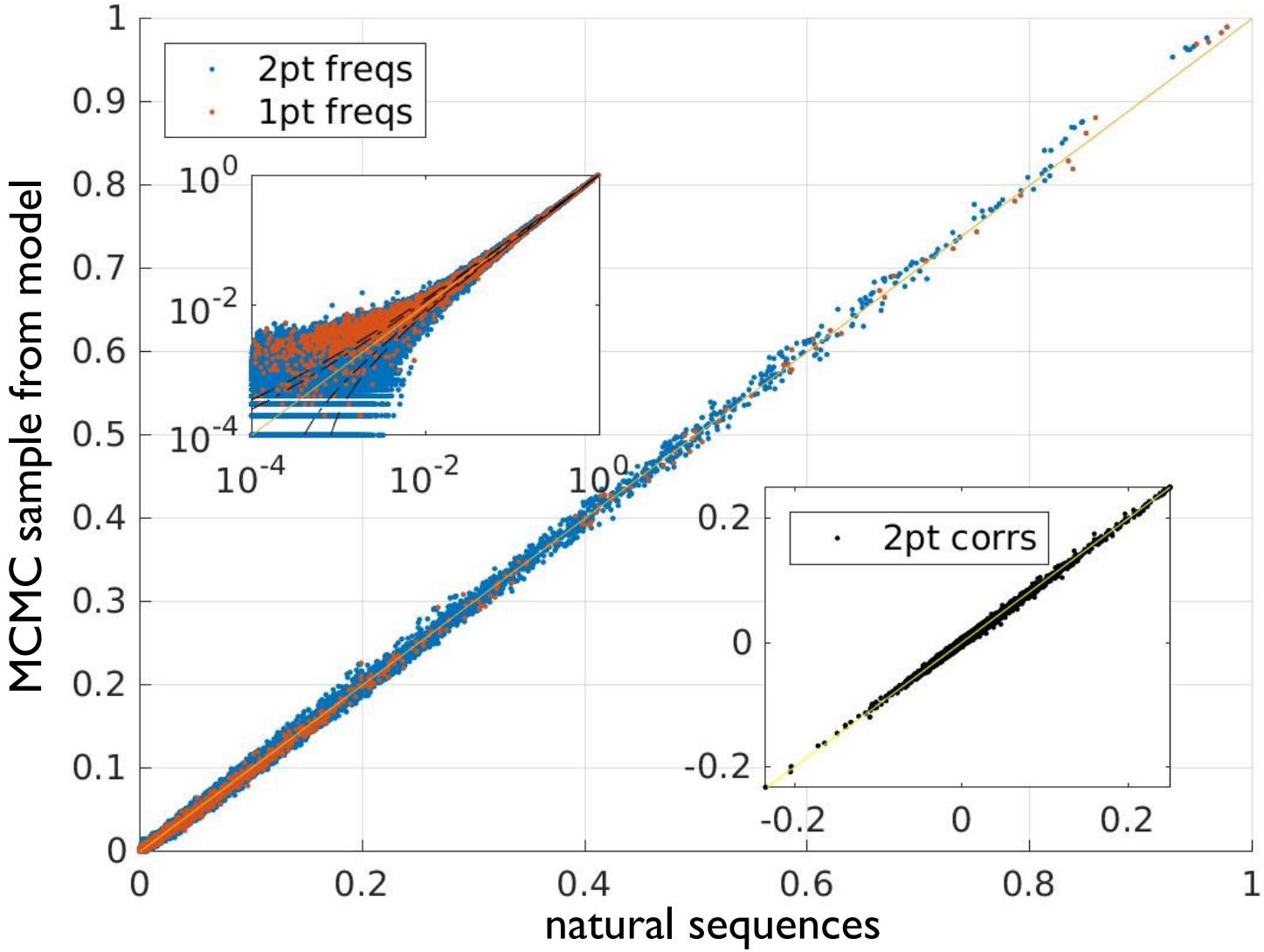
Input data:

$$\{\overline{S}^\mu\}_{\mu=1,\ldots,M}$$



DCA reaches very precise fitting

response regulator (PF00072)

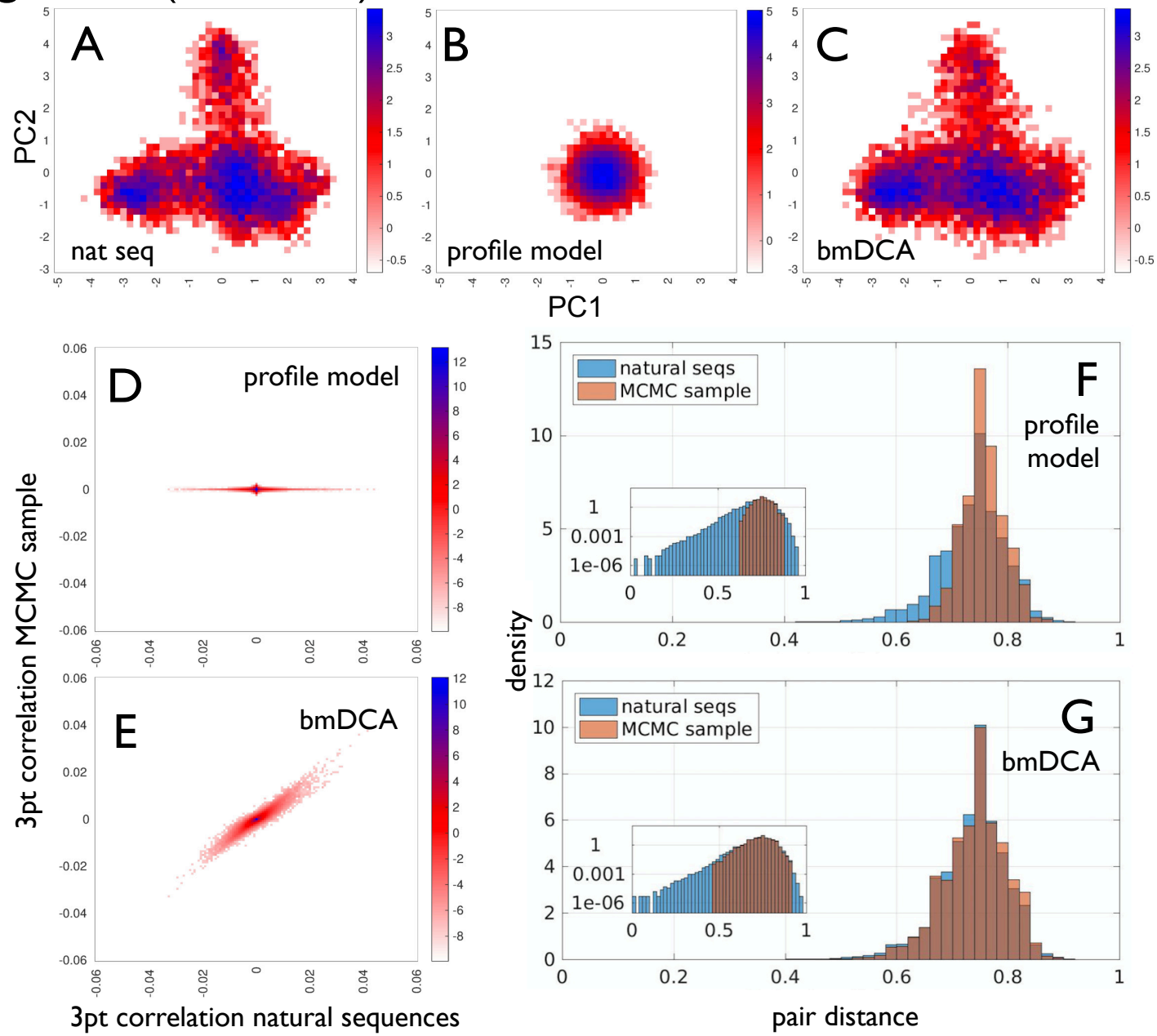


➡ tested for many protein families

[Figliuzzi, Barrat-Charlaix, MW, *Mol. Biol. Evol.* 2018]

DCA reproduces non-fitted quantities

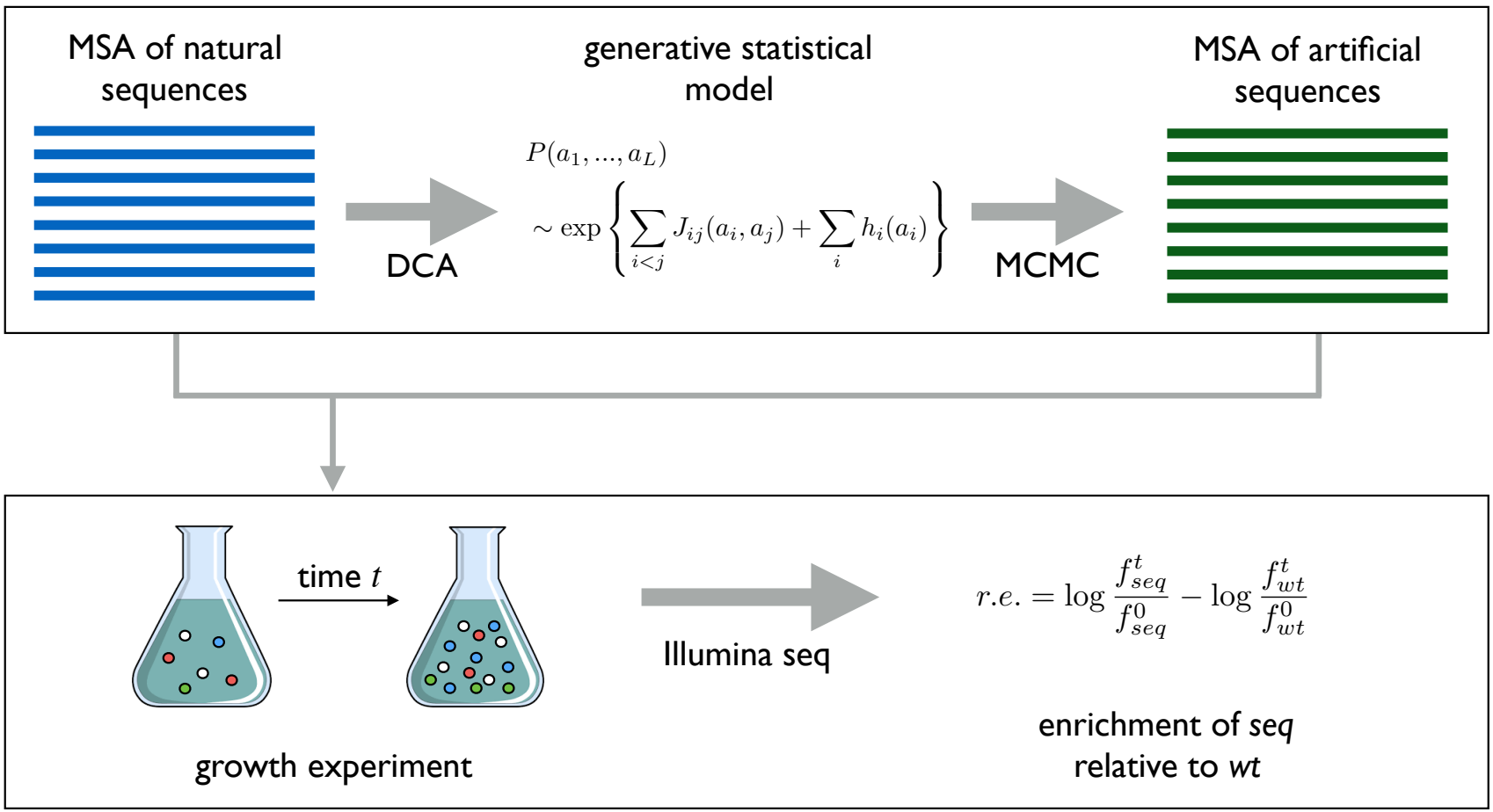
response regulator (PF00072)



➡ tested for many protein families

[Figliuzzi, Barrat-Charlaix, MW, Mol. Biol. Evol. 2018]

Coevolution-guided protein design



data-driven statistical sequence design

experimental validation of natural and designed sequences

experiments by B. Russ, R. Ranganathan

…all models are wrong, but some are useful.

[George E.P. Box, 1976]

Biological sequence evolution

≠

MCMC sampling from Potts models

…all models are wrong, but some are useful.

[George E.P. Box, 1976]

However DCA allows for

- protein structure prediction – strong couplings accurately predict contacts
- mutational effect prediction – statistical model as sequence landscape
- generative statistical modeling – pairwise couplings seem necessary and sufficient to capture biological sequence diversity

…all models are wrong, but some are useful.

[George E.P. Box, 1976]

However DCA allows for

- protein structure prediction – strong couplings accurately predict contacts
- mutational effect prediction – statistical model as sequence landscape
- generative statistical modeling – pairwise couplings seem necessary and sufficient to capture biological sequence diversity

A few current limitations

- many parameters vs. limited data – strong regularization against overfitting
  ➡ parsimonous modeling / dimensional reduction ?
- correlated data – DCA fits functional and phylogenetic correlations
  ➡ inference from non-i.i.d. samples ?
- better models may be even more useful…
  ➡ systematic selection of statistically relevant observables ?
- uses multiple-sequence alignment based on profile models
  ➡ more accurate methods using couplings for alignment ?
- sequences modeled as strings over abstract 21-letter alphabet
  ➡ prior biological knowledge / integrative modeling ?

Thanks to:

The group in Paris:

Juliana Bernardes
Pierre Barrat-Charlaix
Giancarlo Croce
Kai Shimagaki
Edwin Rodriguez
Nika Abdollahi
Anna-Paola Muntoni
Maureen Muscat

Francesco Oteri

Alumni:

Maria Virginia Ruiz Cuevas
Eleonora de Leonardis
Guido Uguzzoni
Alice Coucke
Matteo Figliuzzi
Christoph Feinauer

Collaborators:

Terry Hwa (UC San Diego)
Hendrik Szurmant (Western U LA)
Alexander Schug (KIT Karlsruhe)
Jose Onuchic (Rice U, Austin)
Faruck Morcos (UT Dallas)
Angel E. Dago (Scripps La Jolla)
Joanna Sulkowska (U Warsaw)
Erik Aurell (KTH Stockholm)
Andrea Pagnani (Politecnico Torino)
Thomas Gueudré (IIGM Torino)
Carlo Baldassi (U Bocconi Milano)
Rémi Monasson (ENS Paris)
Simona Cocco (ENS Paris)
Olivier Tenaillon (Inserm Paris)
Bill Russ (UTSW Dallas)
Rama Ranganathan (U Chicago)
Anne-Florence Bitbol (Sorbonne U Paris)
Francesco Zamponi (ENS Paris)

Funding:

Agence Nationale de la Recherche ANR
îledeFrance Demain s'invente ici
ICS Institut du calcul et de la simulation
Horizon 2020 European Union funding for Research & Innovation