

Data-driven models of sequence landscapes

Martin Weigt

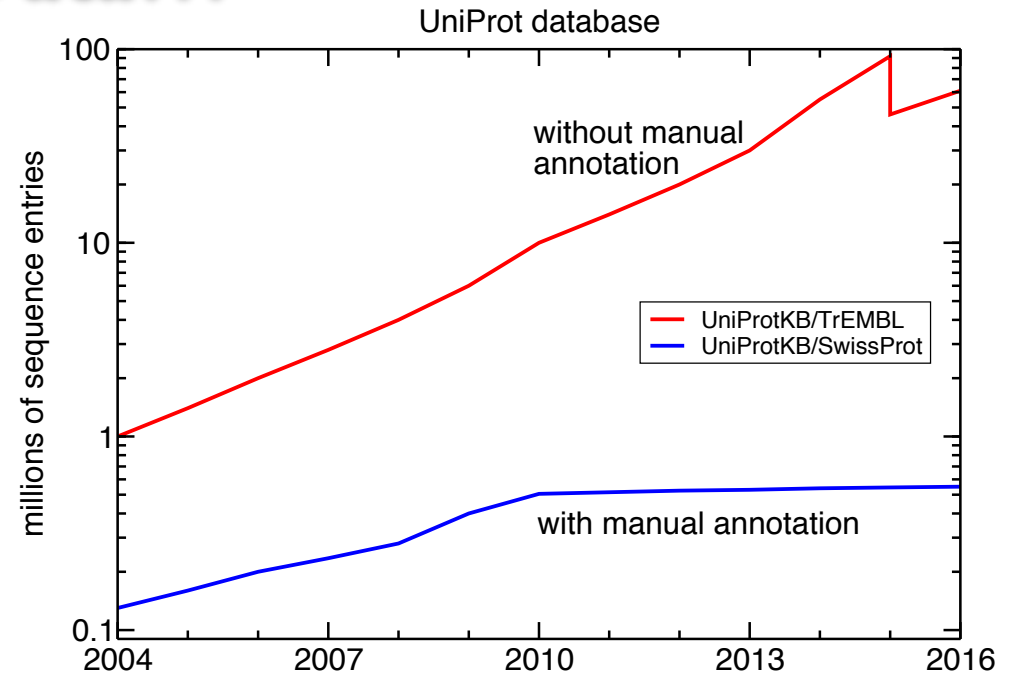
Laboratoire de Biologie Computationnelle et Quantitative
Sorbonne Université

All models are wrong, but some are useful.

[George E.P. Box, 1976]

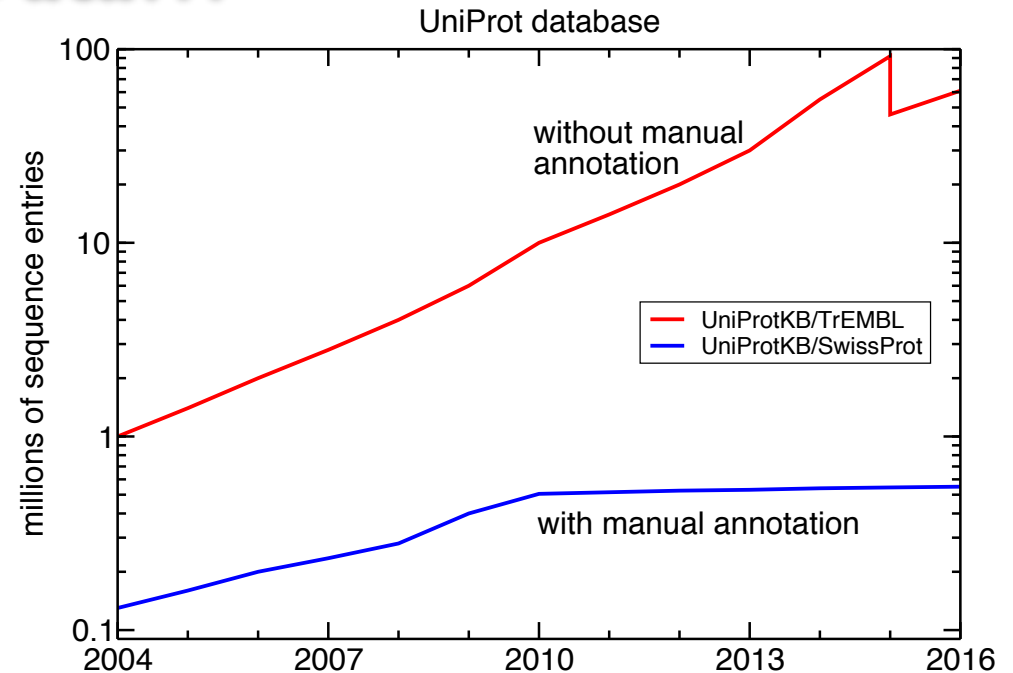
Data...

Sequence data are rapidly accumulating...



Data...

Sequence data are rapidly accumulating...



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT



Pfam 30.0 (June 2016, 16306 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	<input type="text" value="enter any accession or ID"/> <input type="button" value="Go"/> <input type="button" value="Example"/>

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

...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 :

```
-LNQFADDLAHELRTPVNILLGKNQVMLS-QERSAEEYQQALVDNIEELEGSLRTENILFLARAEH-  
ALGELTAGIAHEINNPTAVILGNTELIRFLGADASRV-EEEIDAILLQIERIRNITRSLQYSRQ--  
SQRQFVTNASHELKTPIAIISANTEVLEI-----TMGK-NQWTETILKQVKRLSGLVNDMVALAKLEE-  
---AFVSNASHHELRTPVTSIKGFAETIKG-MSAEFEAKDDFLDIIYKESLRLEHIVEHLLTLSKAQ--  
-VGQLTGGIAHDFNNMLTGVIGSLDLIKLS-----GRLVERFMDAALISAQRAASLTDRLLAFSRRQS-  
---RMTHQVSHVEGNMIGIITGSLGLLERETGFNDRQ-KRHIARIRKAADRGRSLASSMLTIGS----  
ALGEMLDHIAHQWKOPINSISLTAQDMADYGELTDGDVQTTIDKIMSLLHMSQTVDFRGGFYR----  
-VGRLAGGVAHDFNNLLSVINGYCEMLAA-QVSDRPQALREVSEIHRAGLRAAGLTRQLLAFGRRQ--  
SLGELAAGVAHEINNPNAVILLNVDLVKKWSEMSEEL-PLLLTEMEEGAGRIKRIVDDLKDFARGD--  
-MGEFAAYIAHEINQPLSAIMTNANAGTRNPSNIPEAKEALARIIRDSDRAAEIIRMVRSFLKRQ--  
---GQLAGGIAHDFNNILQIISGNTQILQYQTNPDPP-----QLLEILKAVERGTALTRSMALAFSRKQT-  
---GQLTGGIAHDFNNLLQVILGNLEFVRAKLDGDAK-LQTRIERAAWAAQRGATLTGQLLAFARKQ--  
AKTDFLSNMSHEIRTPLNAILGFIQVLKD-AEMKPKD-REYLELMDESSKNLLSLVNDIIEIDLIESG  
--GREVLHLVHDLKTPLATIEGLVSLMET-RWPDPKM-QEYCQTIYGSITSMSKMVSEILY-----  
-RARLLADVAHELRTPVATLTGYLEAVEDVRPLDAST-----IAVLRDQAVRLTRLAQDLADVTHAEGG  
SMKRMLTNMSHDLKTPLTVILGYIETIQSDPNMPDEERERLLGKLRQKTNELIQMINSFFDLAKLES-  
AKSEFLANMSHELRTPLNAIIGFSEMIQAFGPLGSDRYEEYINDIHTSGNFLNVINDILDMSKIEAG  
-MQRFIADATHQLRTPLAAIDAEEVELLTD-QTRDPKA----LDKLRGRIADLARLASQLLDHAM----  
-RKKAVHTITHELRTPLTAITGYAGLIRK-EQCEDKS-GQYIQNILQSSDRMRDMLNTLLDFFRLDNG  
-REEFMNMTSHELMNPLSAAVQAHTMISLHDDNSKSNIEIAKIIACGEHQKLVEDARMMSKLD--  
-KSRYVVGLSHELRSPLNAISGYAQLLEQDTSLAPKP-RDQVRVRRSADHLSGLIDGILDISKIEAG  
----AFSYMRHAINNPLSGMLYSRKALKN-TDLNEEQ-MRQIHVSDNCHHQLNKILADL-----  
-QENFIDMTSHEMRNPLSAILQCSDEITST-----LCLEAANTIALCASHQKRIVDDILTFSKLDL-  
SQRTLNAIAHDLRQPLYRIRFALEMFND-SLLSIEQRQQYRQSIENSLRDLHDHINQSLQLSRYT--  
---KLLLLSLSHDIKTPLSAIKLNKALSRLYKDAEKQ-REAAEHINARADEIENFVSRITKASSE---  
---HAFIADAHELRTPLTALKLQLQLTER---ATSDVREVGFKLNERLDRSIHLVKQLLTLARSES-  
-QKNFISNASHELNTPLTSIIVTADLALS-KQRTDEEYRTALSRIIMDAAGHLE-----  
-RGALLTSISHDLRTPLASILGATSSLESGEELDENARKELLSTIHDEADRLNRFVANLLDMTRLEAG  
-KSEFLANMSHELRTPLNGVIGFTRTLK-TELTPTQ-RDHLNTIERSANNLLAIINDVLDVDFSKLEAG  
AKSEFLANMSHDIRTPMNAITGMTAIAATA-HIDDPKQVKNCLRKIALSSRHLLGLINDVLDMSKIESG  
-LSQFSADLAHDFRTPLANLIGQTEVTLA-HPRSAAEYRAVLESSLEEYARLSRMIEDMLFLARADH-  
SKSMFLATVSHELRTPLYGIIGNLDLLQT-KELPKGV-DRLVTAMNNSSSLLKIIISDILDFSKIES-  
AKTAFLATLSHEIRTPMNGVLGTAQILLK-TPLSTEQ-EKHLKSLYDSGDHMMTLLNEILDVDFSKIEQ  
SKKQLIDGIAHELRTPLVRLRYRLEMSN---LTPPE---SQALNRDIGQLEALIEELLTYARLDR-  
-KTQFFINTAHDIRTPLTIKAPLEELLEEEETLTDNG-ITRTNIALRNVEVLLRLVSNLINFERT---  
---VFIDNMTHEMKTPLTSIIGFSDLLRS-ARLDDETVDHYAESIYKEGKYLKSISSKLMDL-----
```

multiple-sequence alignment

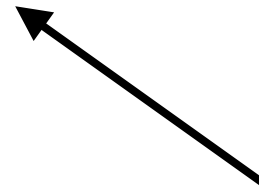
- 50-500 positions
sequence length
- 10^3 - 10^5 homologous
sequences

...

Inference of statistical sequence models

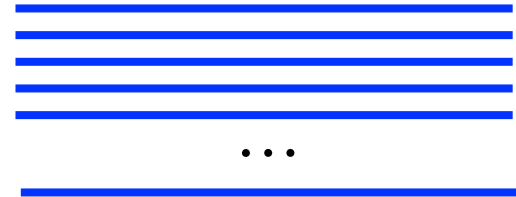
From data to statistical models

$$P(\bar{A}) = P(A_1, \dots, A_L)$$



Data – MSA

$$\{\bar{A}^\mu\}_{\mu=1\dots M}$$



Attention:

Data insufficient to do this without modeling

Inference of statistical sequence models

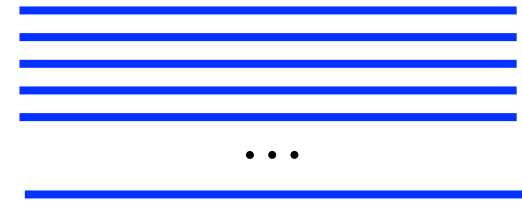
From data over observables to statistical models

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

maximum
entropy
model

Data – MSA

$$\{\bar{A}^\mu\}_{\mu=1\dots M}$$



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

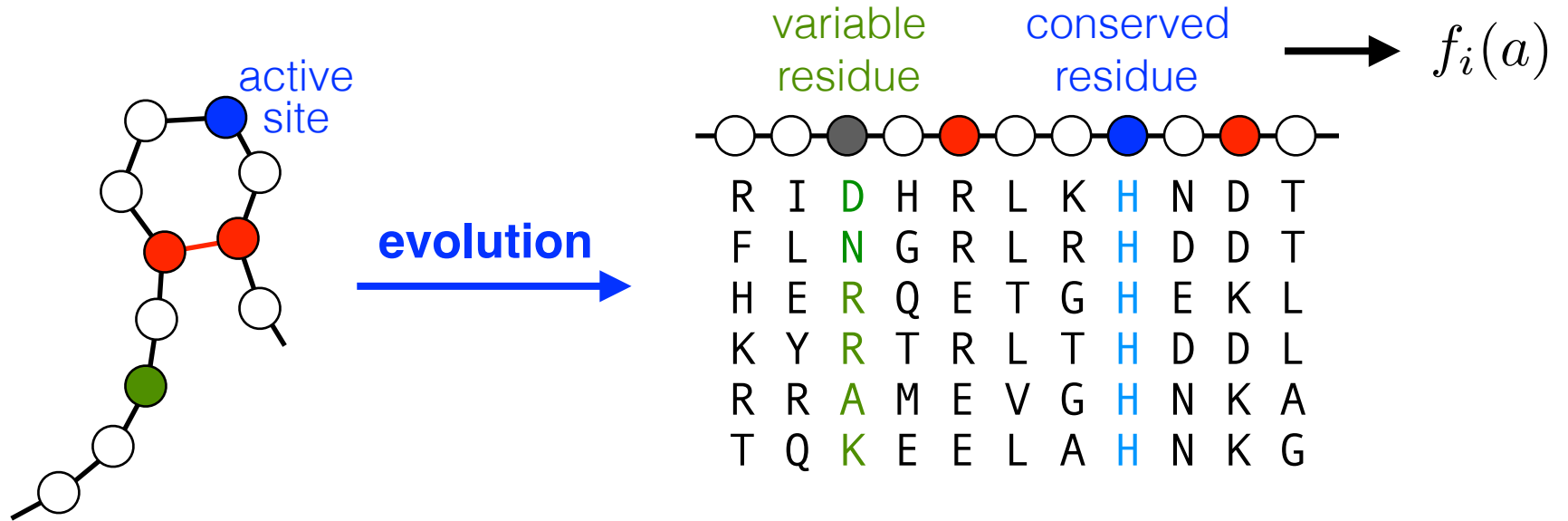
model average

average over data

Attention:

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

Conservation in proteins

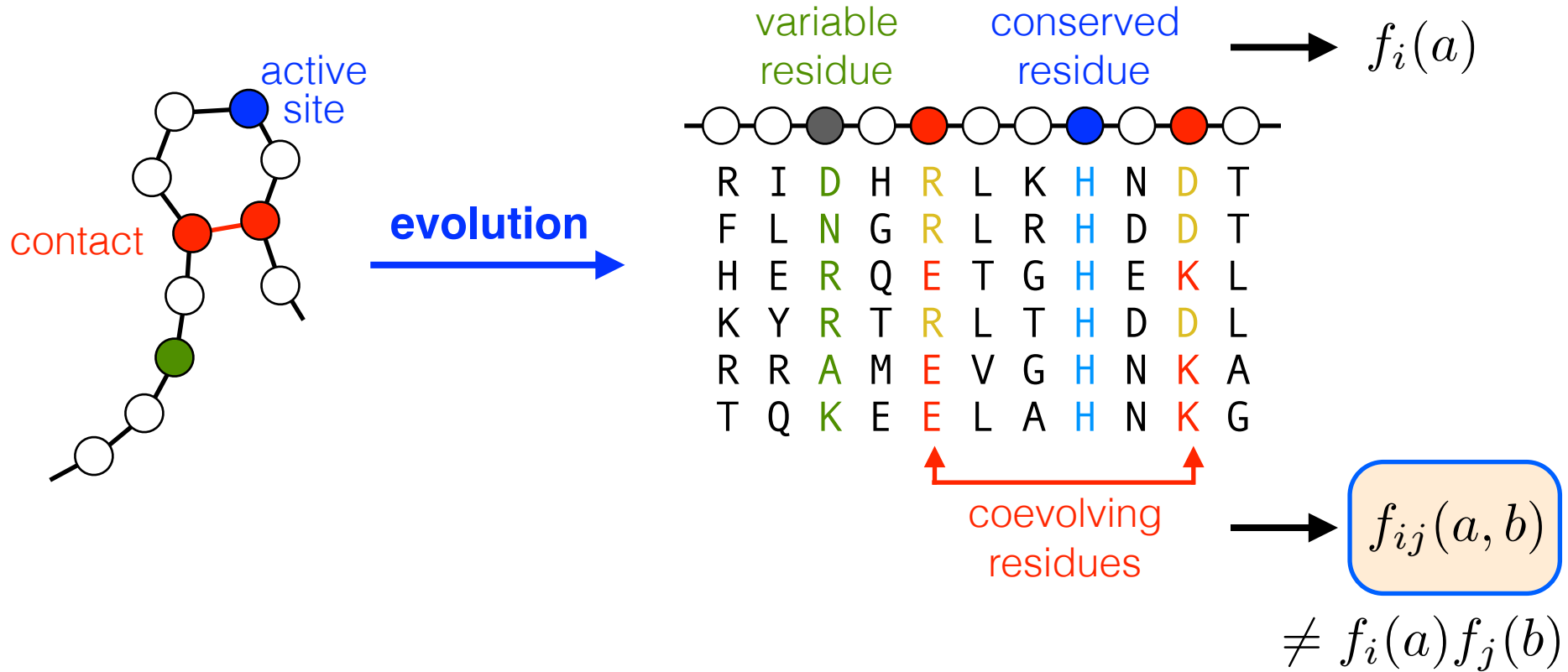


Profile model – independent residues

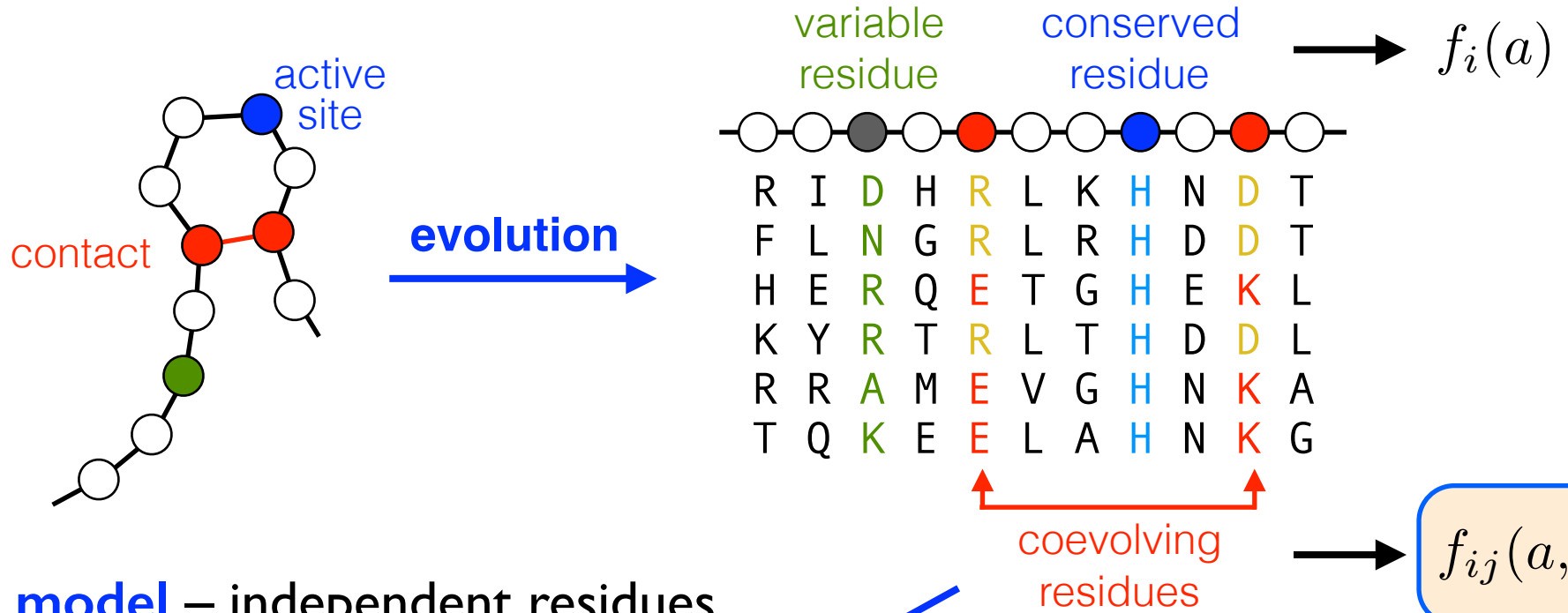
$$P(a_1, \dots, a_L) \sim \exp \left\{ \sum_i h_i(a_i) \right\}$$

statistical modeling

Conservation and coevolution in proteins



Conservation and coevolution in proteins



Profile model – independent residues

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

statistical modeling

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

$$P(a_1, \dots, a_L) \sim \exp \left\{ \sum_{i < j} J_{ij}(a_i, a_j) + \sum_i h_i(a_i) \right\}$$

[Weigt et al, PNAS '09]
[Morcos et al, PNAS '11]

Direct coupling analysis (DCA)

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

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

- 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 $\sim 2^L$
 - ➔ approximations (MCMC) needed for computational efficiency

Direct coupling analysis (DCA)

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

$$\begin{aligned} P_1(A_1) &= \sum_{\{A_i | i > 1\}} P(A_1, \dots, A_L) \\ &= \sum_{\{A_i | i > 1\}} P(A_1 | A_2, \dots, A_L) P(A_2, \dots, A_L) \\ &\simeq \frac{1}{M} \sum_{m=1}^M P(A_1 | A_2^m, \dots, A_L^m) \end{aligned}$$

biological sequence data



[Balakrishnan et al., Proteins '11]

[Ekeberg, Lövkvist, Lan, MW, Aurell PRE '13]

Are pairwise DCA couplings useful?

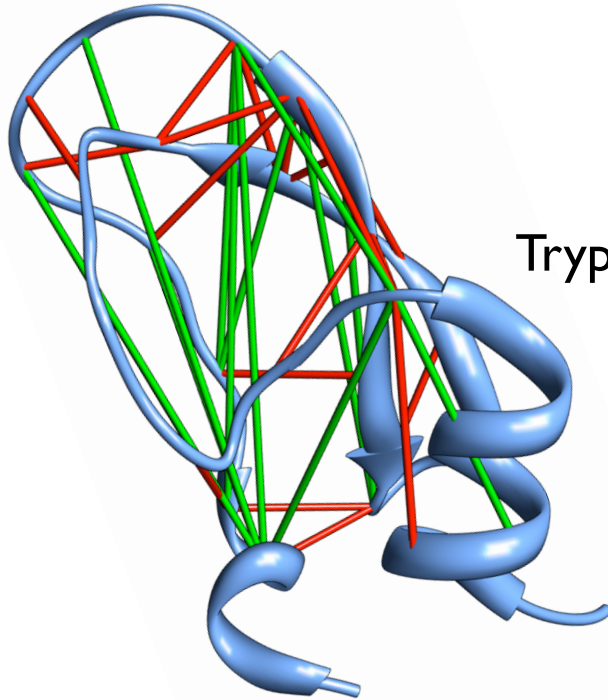
DCA models are **graphical statistical models**:

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

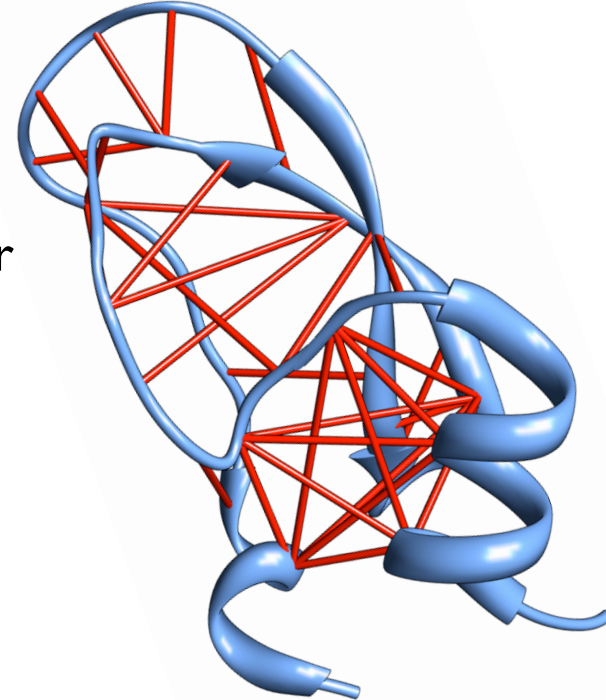
What is the **biological information** contained in such models ?

Strong couplings predict residue contacts

30 strongest **correlations**



30 strongest **couplings**



Trypsin inhibitor

- 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]

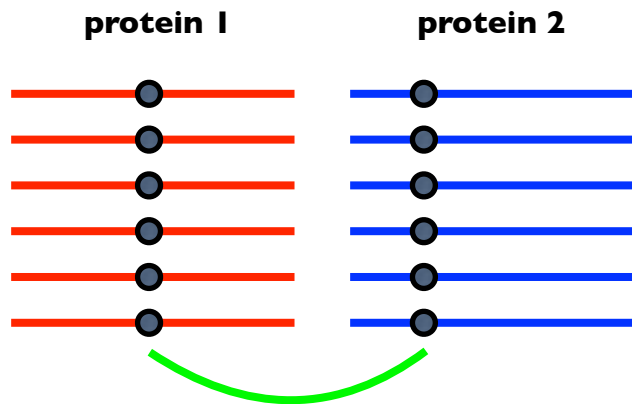
[Nugent, Jones, PNAS '12]

[Ovchinnikov et al., eLife '15]

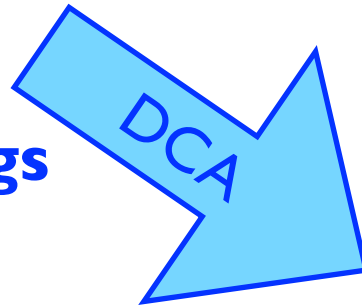
[Ovchinnikov et al., Science '17]

Prediction of inter-protein residue coevolution

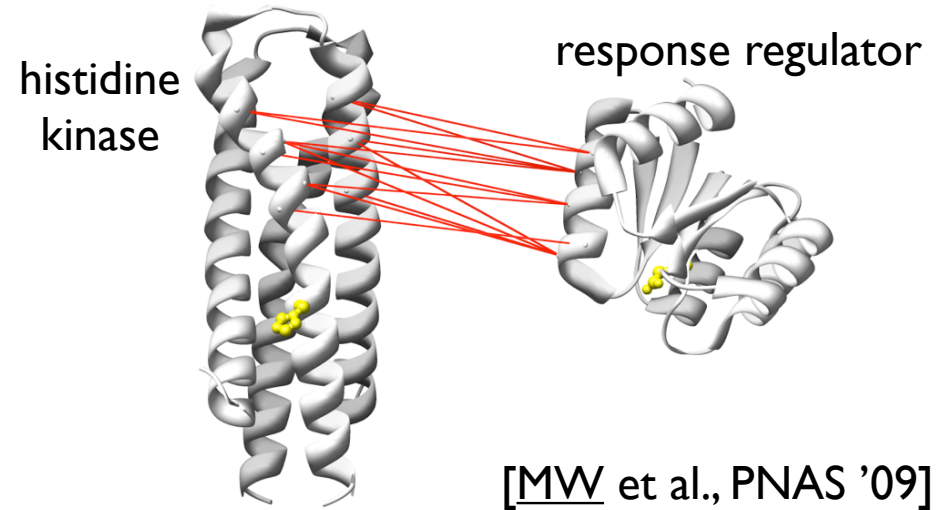
Joint MSA of protein families



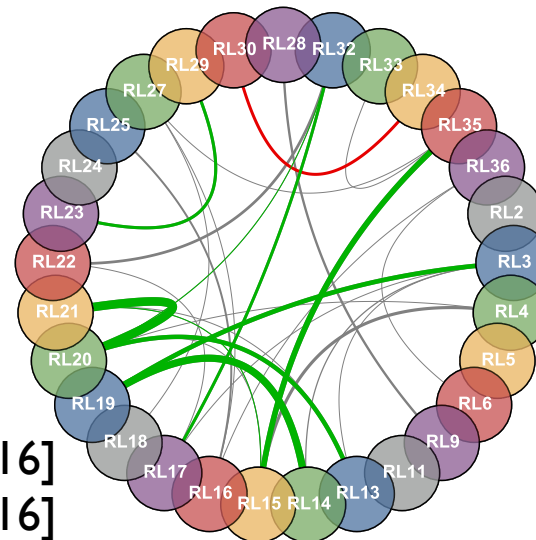
inter-protein couplings



inter-protein residue contacts



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 Barnaudⁱ, Pierre-Alexis Gros^{a,b,j}, and Olivier Tenaillon^{a,b,1}

Deep mutational scanning of proteins

TEM-1 protein

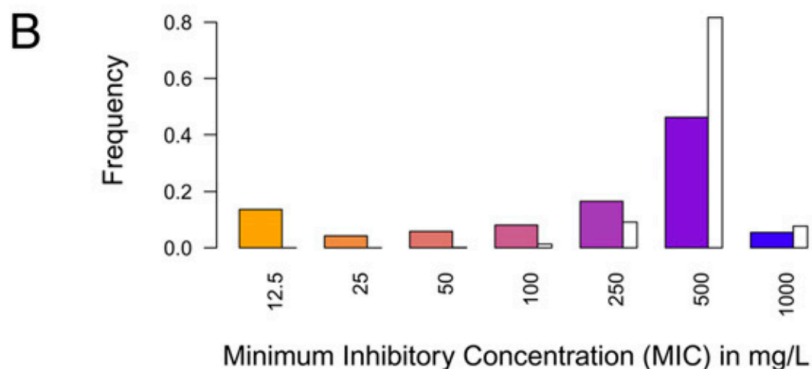
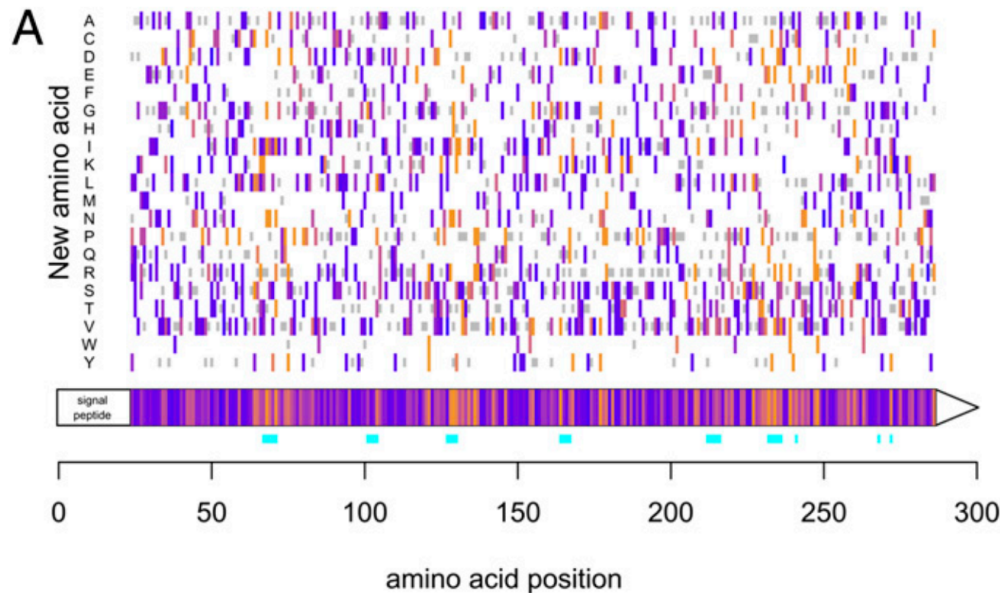
- causes antibiotic resistance

generated $\sim 10^4$ random mutants

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

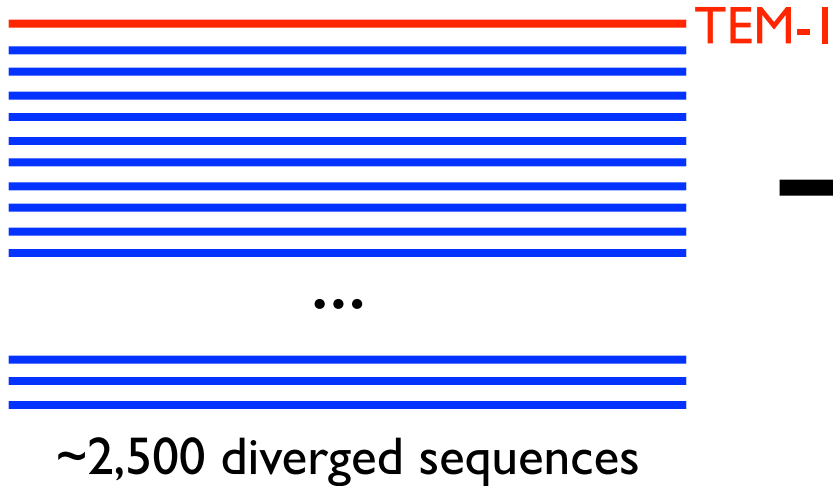
measured resistance to amoxicillin

- minimum inhibitory concentration as proxy for fitness



Landscape inference by Direct-Coupling Analysis

Beta-lactamase2 family (PF13354)



Statistical landscape inference (DCA)

$$P(A_1, \dots, A_L)$$

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



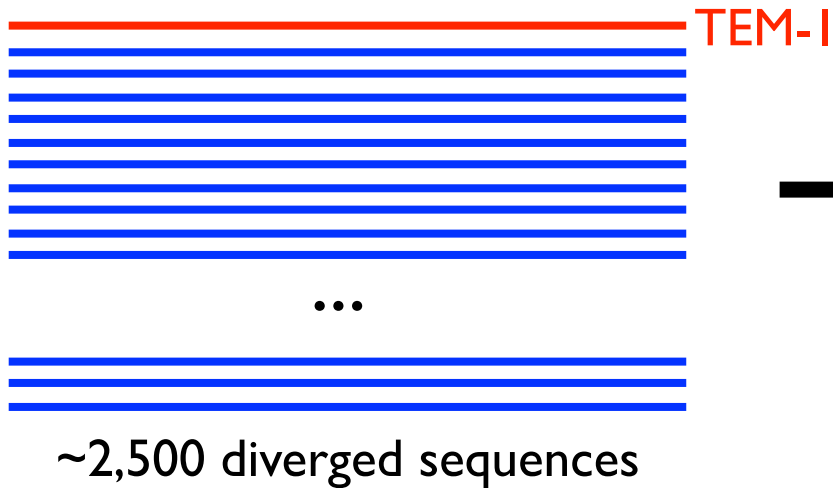
Score for mutant AA sequences

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

Evolutionary constraints
across diverged homologs

Landscape inference by Direct-Coupling Analysis

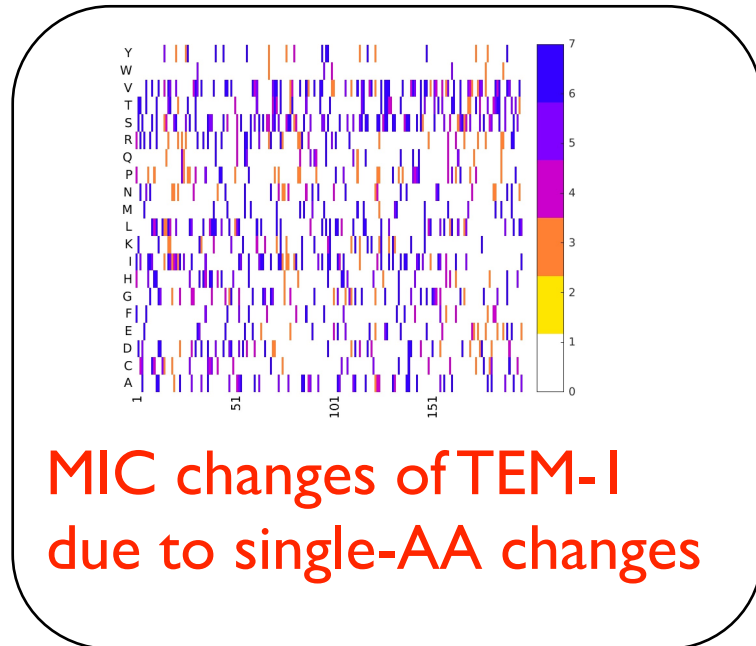
Beta-lactamase2 family (PF13354)



Statistical landscape inference (DCA)

$$P(A_1, \dots, A_L)$$

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



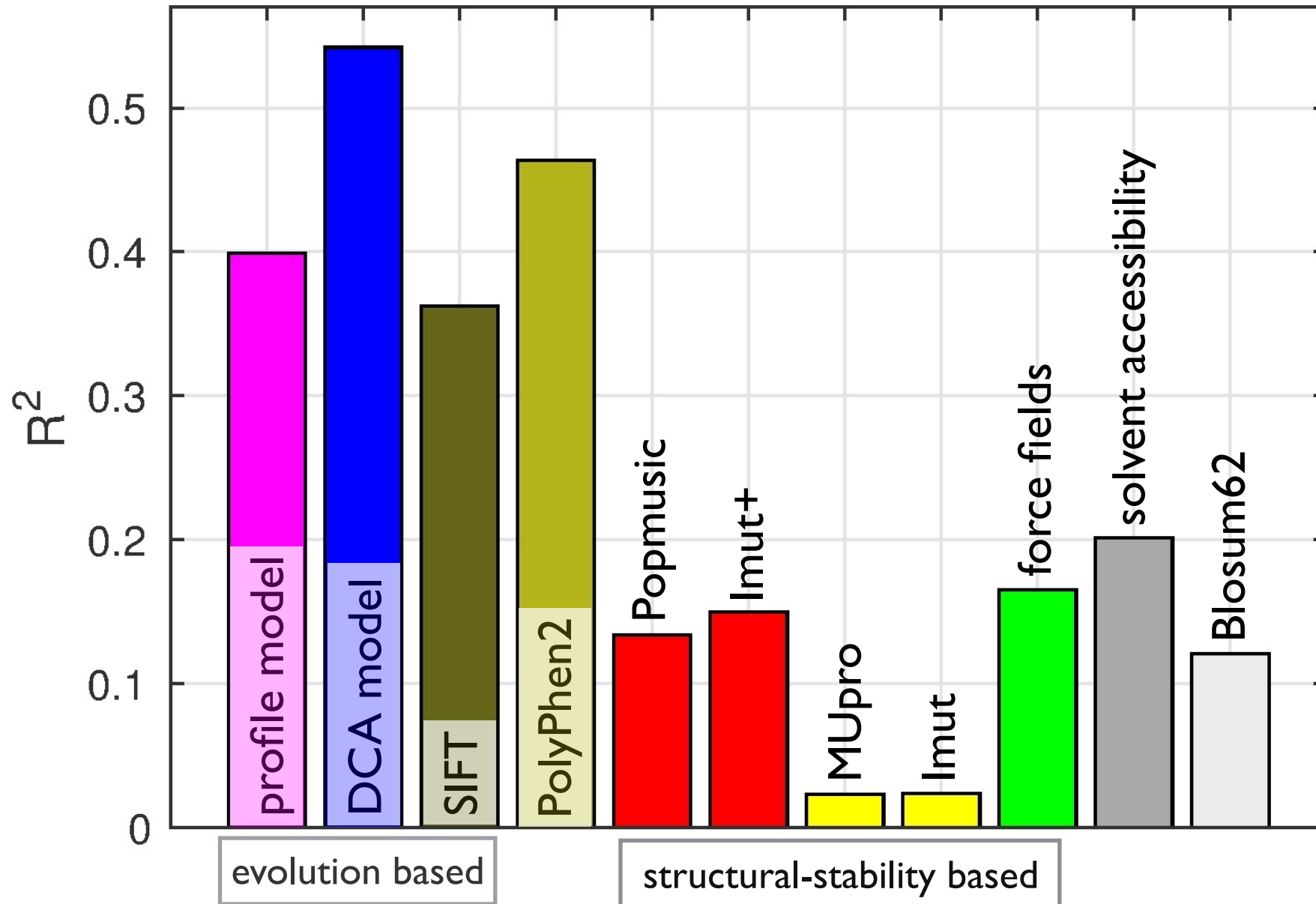
Score for mutant AA sequences

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

Evolutionary constraints
across diverged homologs

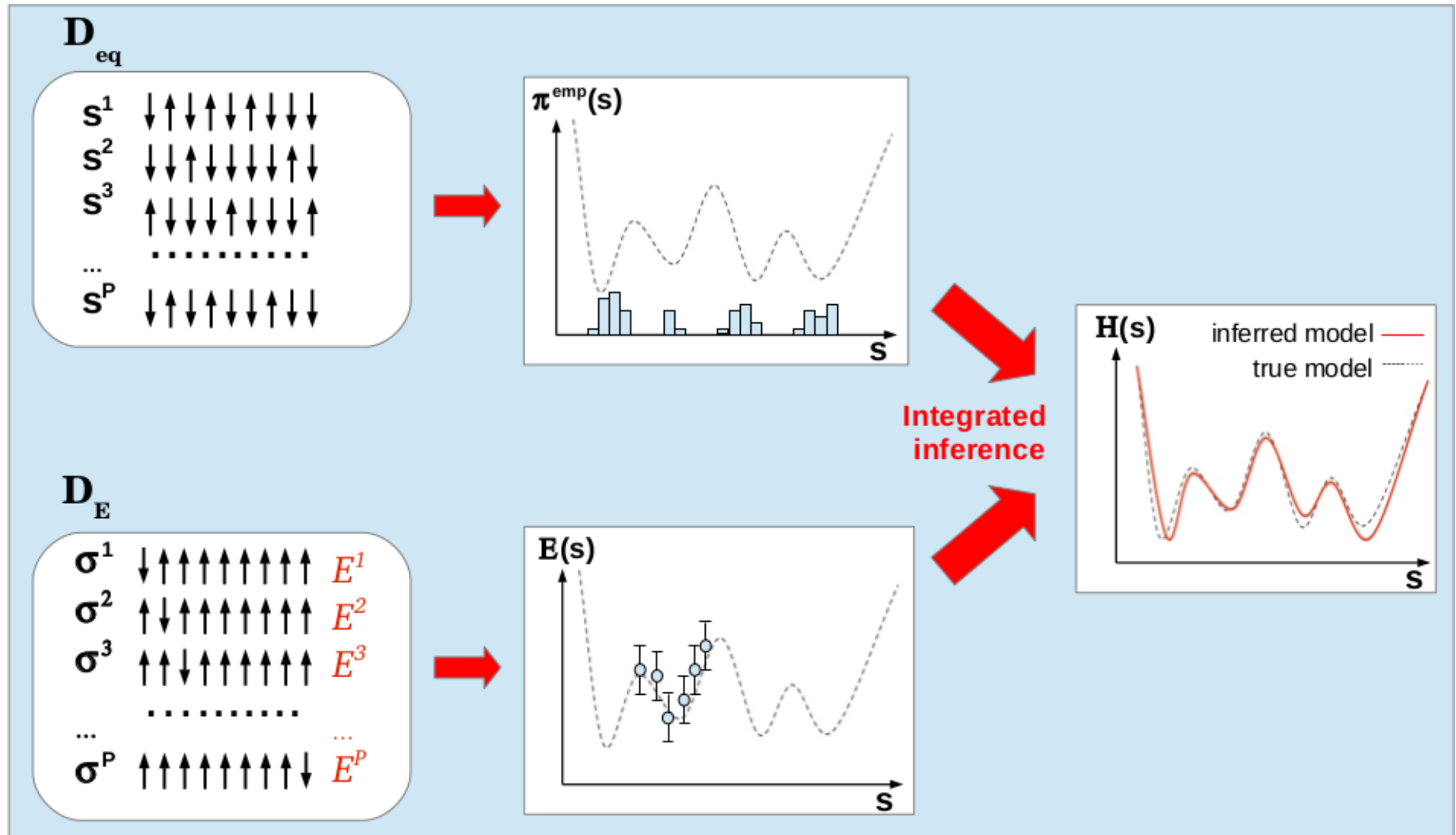


Predicting mutational effects in proteins



Integrating heterogeneous data in landscape inference

MSA: sequence data



DMS: mutational data

Integrating heterogeneous data in landscape inference

Statistical model

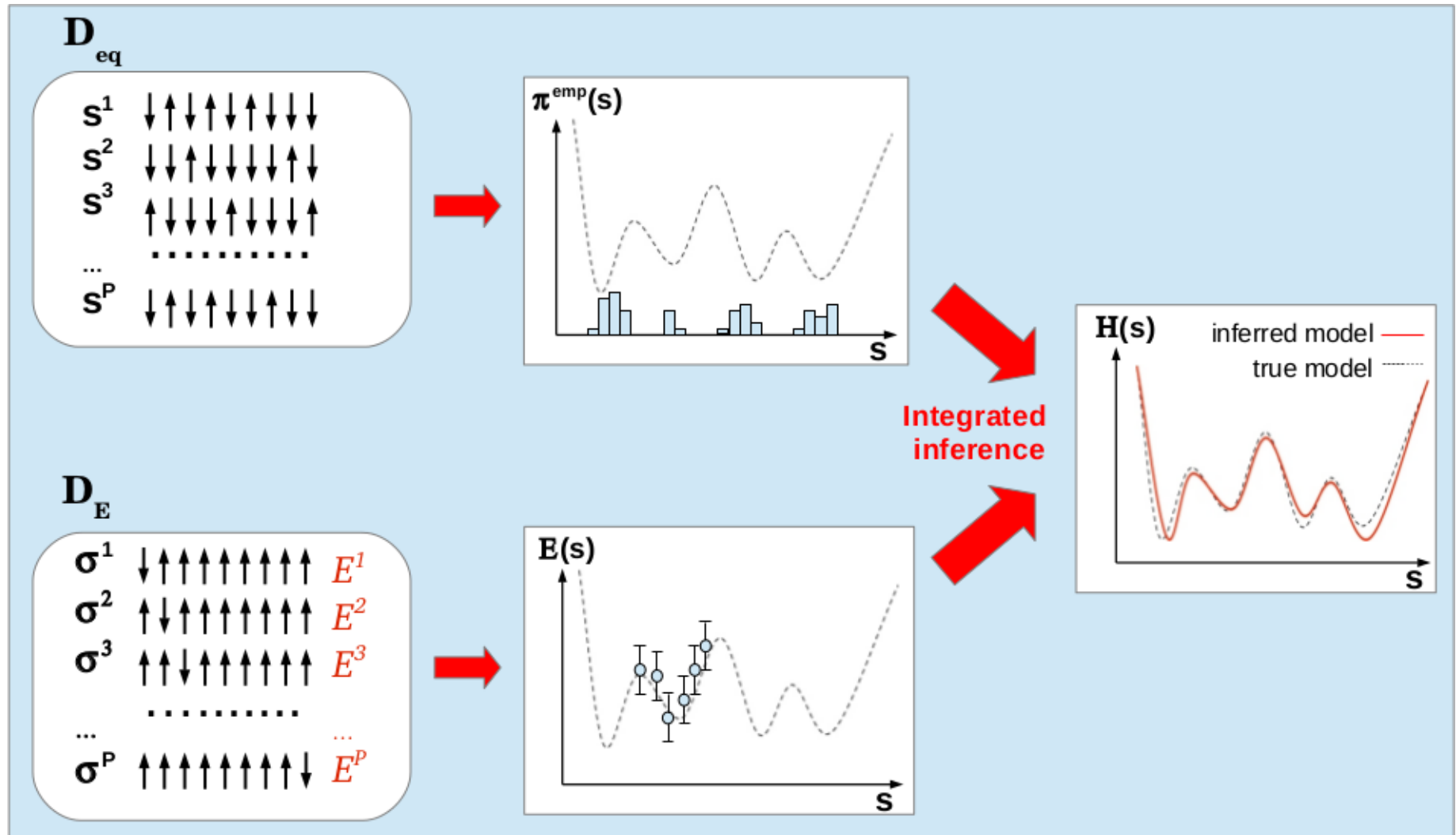
$$P(\bar{A}|\mathbf{J}, \mathbf{h}) = \frac{1}{Z} \exp \left\{ \sum_{a < b} J_{ij}(A_i, A_j) + \sum_a h_i(A_i) \right\} = \frac{1}{Z} \exp \{ -\mathcal{H}(\bar{A}) \}$$

Probability of MSA

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

Integrating heterogeneous data in landscape inference

MSA: sequence data



DMS: mutational data

Integrating heterogeneous data in landscape inference

Statistical model

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

Probability of MSA

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

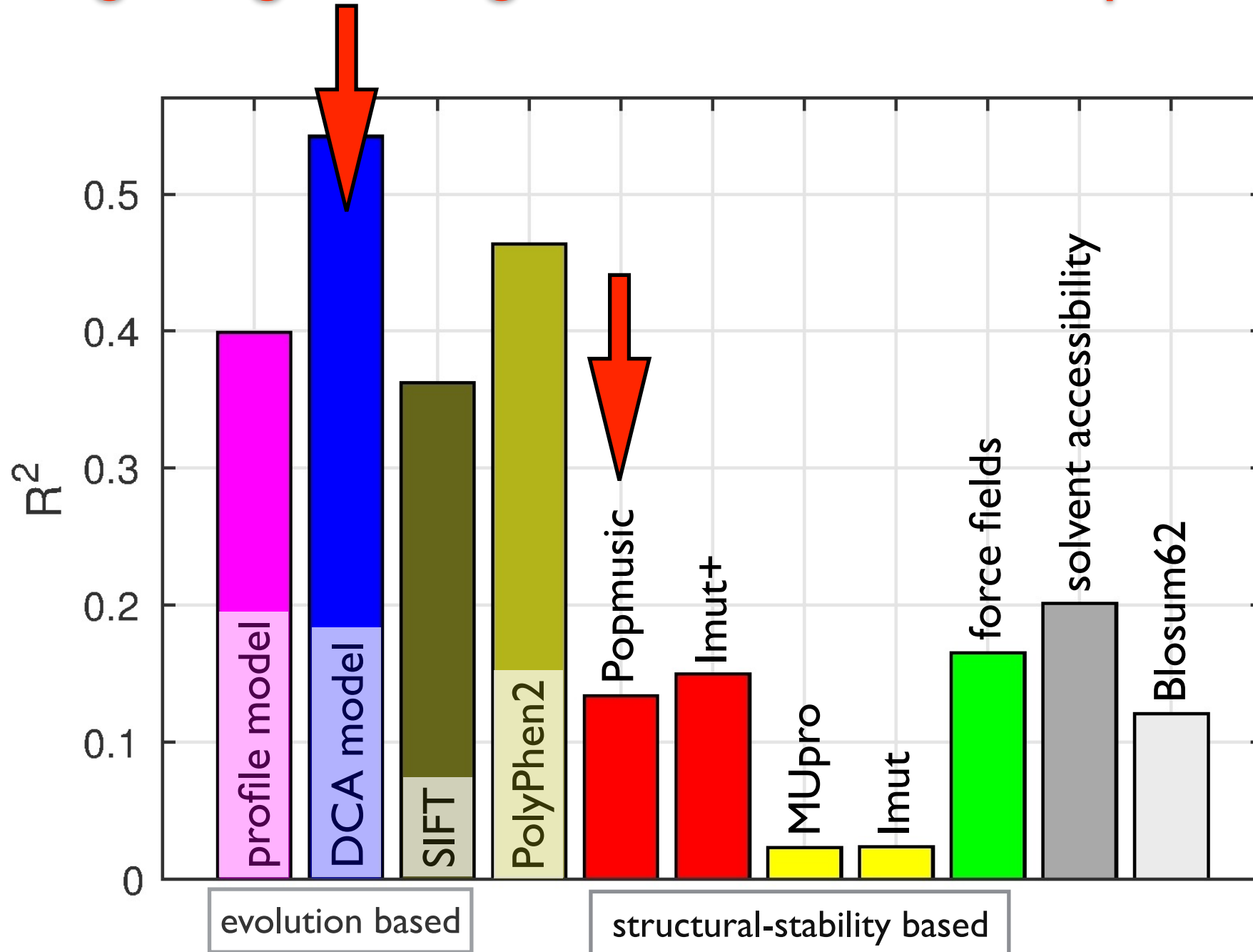
Mutational data - Gaussian experimental noise

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

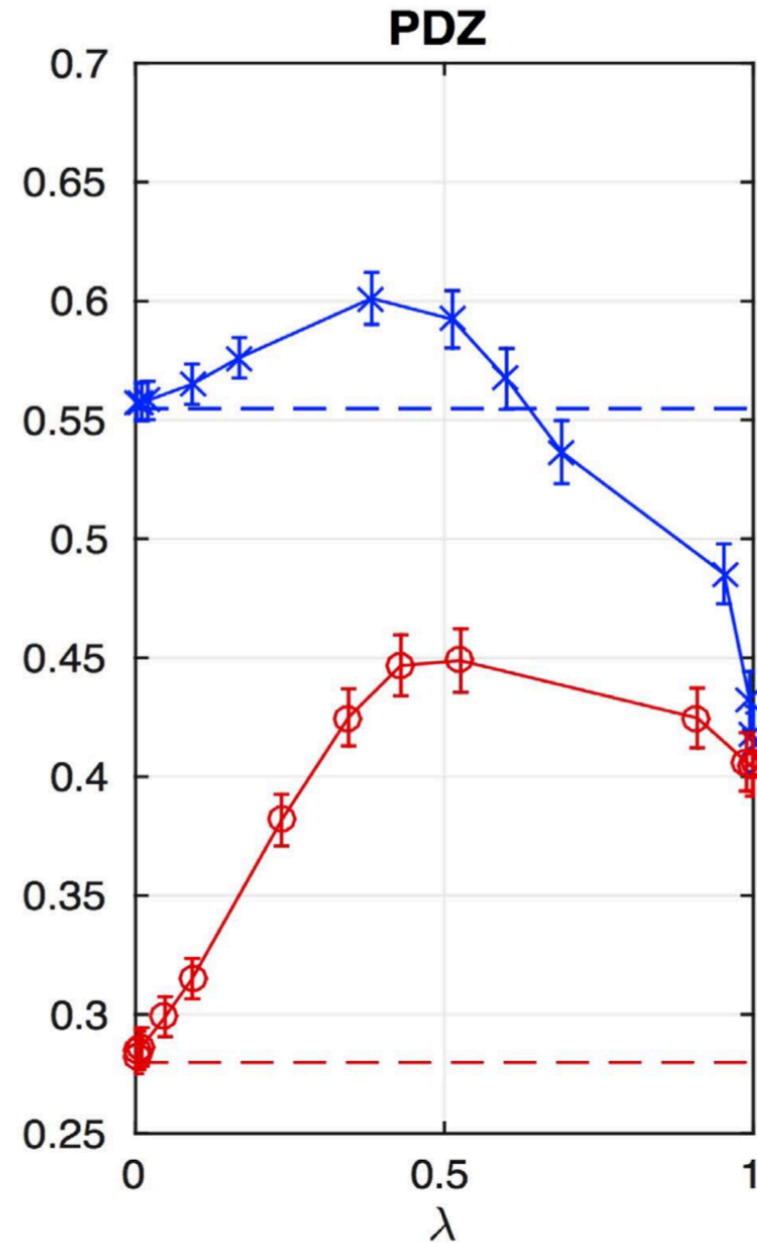
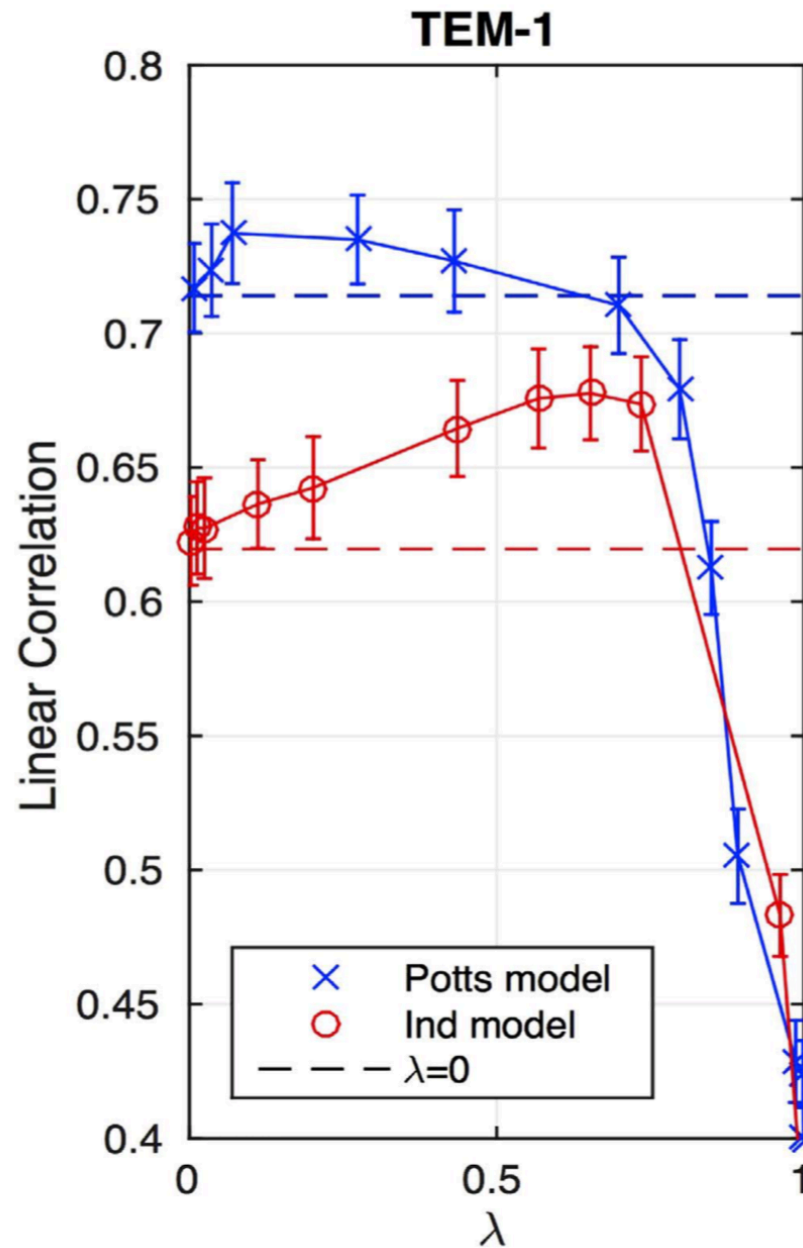
Log-likelihood of model parameters

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

Integrating heterogenous data in landscape inference



Integrating heterogeneous data in landscape inference



Are pairwise DCA couplings useful?

DCA models are **graphical statistical models**:

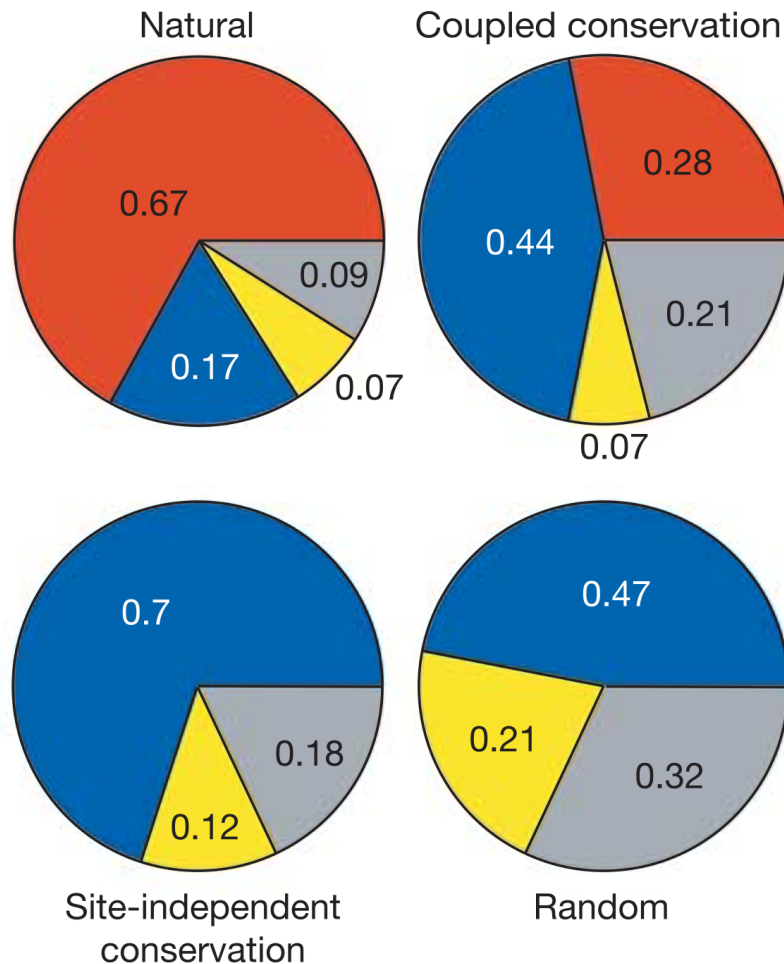
- defined on a **network** of strongly coupled residues
- provide a **probability** to each sequence (sequence landscape)

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

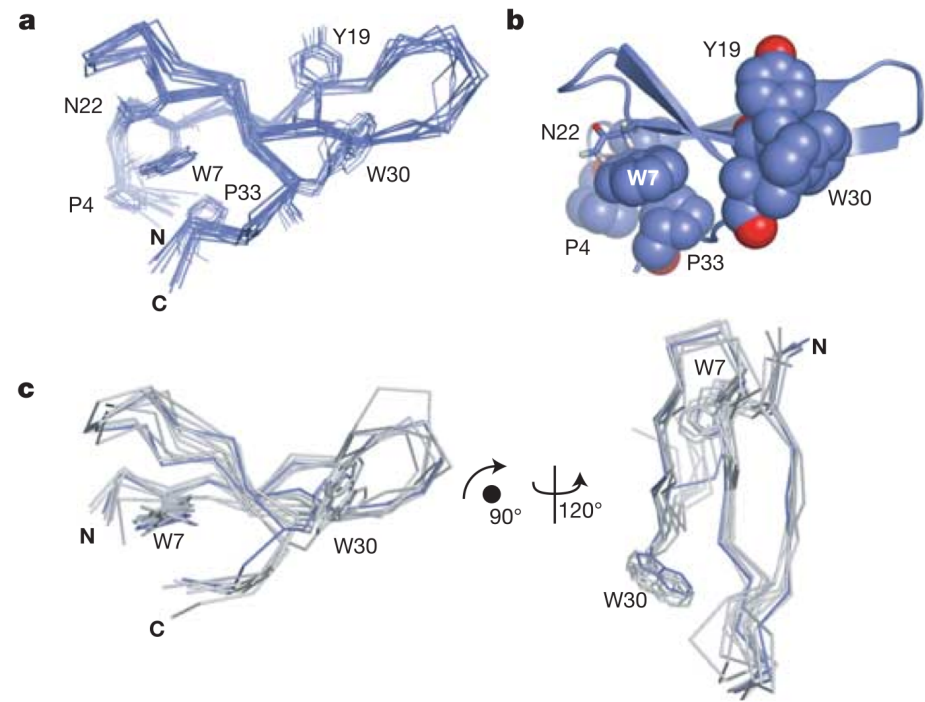
Are pairwise couplings sufficient?

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

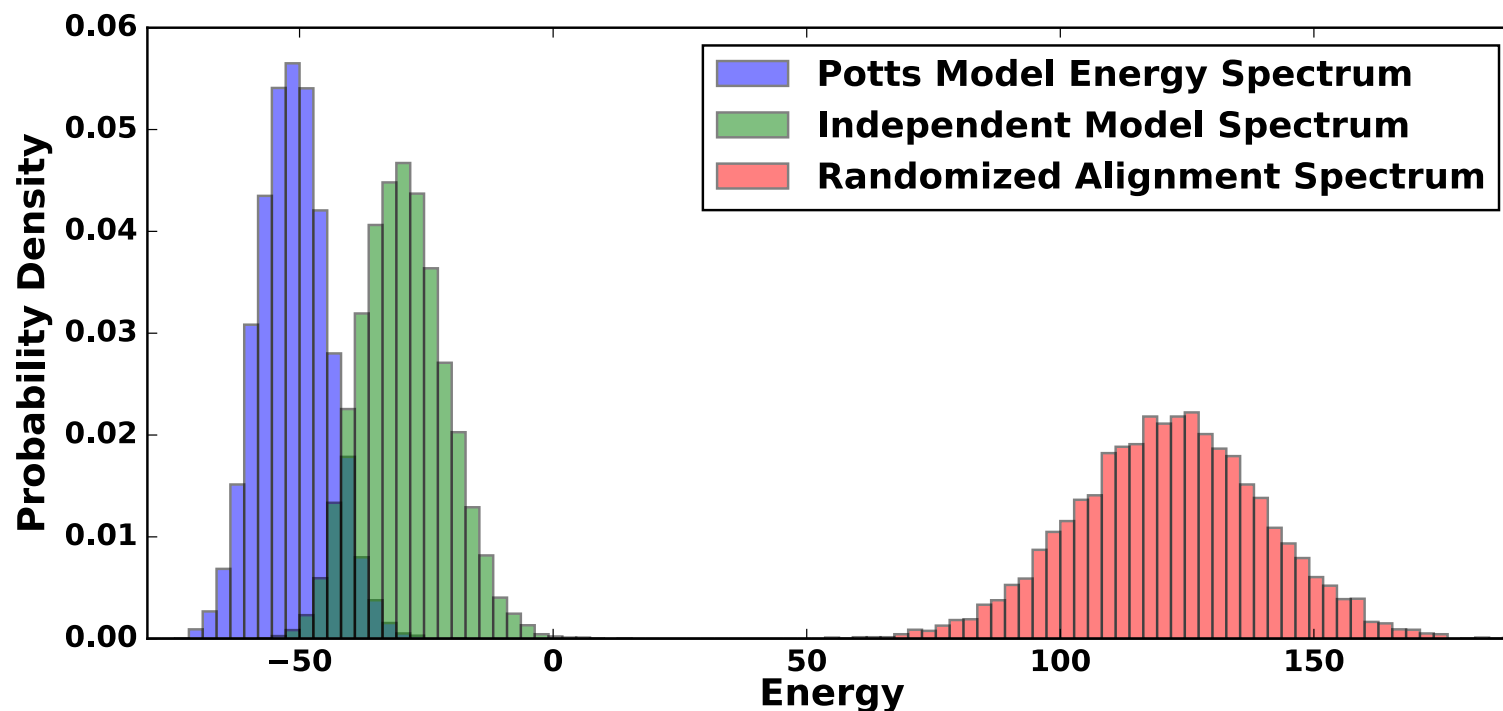


NMR structure of artificial protein (WW domain)



structural alignment with 6 natural WW domains

Predicting folding sequences of the WW domain



high
probability ←

folding

non folding

→ low
probability

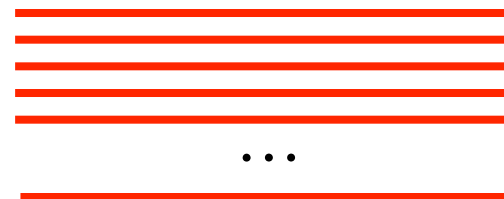
Data-driven statistical design

From data over observables and models to data

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

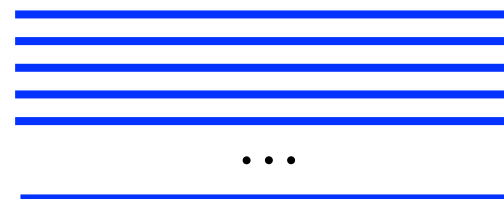
Output data:

$$\{\bar{X}^\nu\}_{\nu=1,\dots,N}$$



Input data:

$$\{\bar{S}^\mu\}_{\mu=1,\dots,M}$$



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



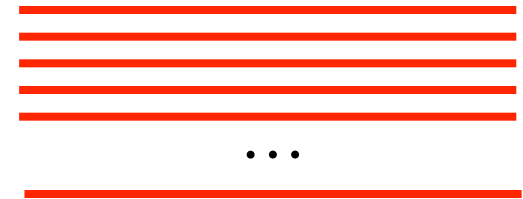
Data-driven statistical design

From data over observables and models to data

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

Output data:

$$\{\bar{X}^\nu\}_{\nu=1,\dots,N}$$

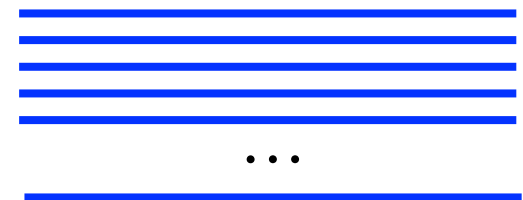


generative model?

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

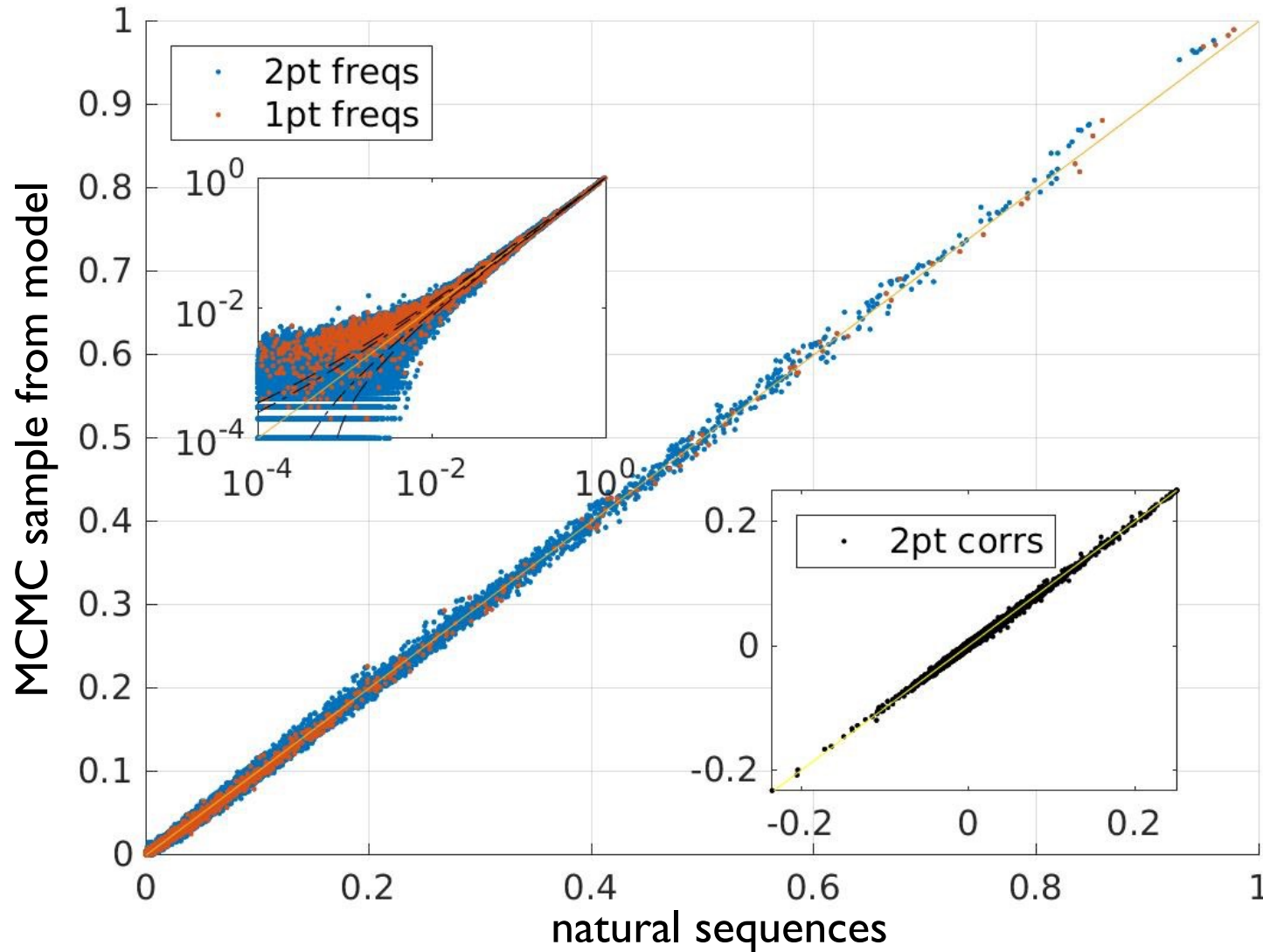
Input data:

$$\{\bar{S}^\mu\}_{\mu=1,\dots,M}$$



DCA reaches very precise fitting

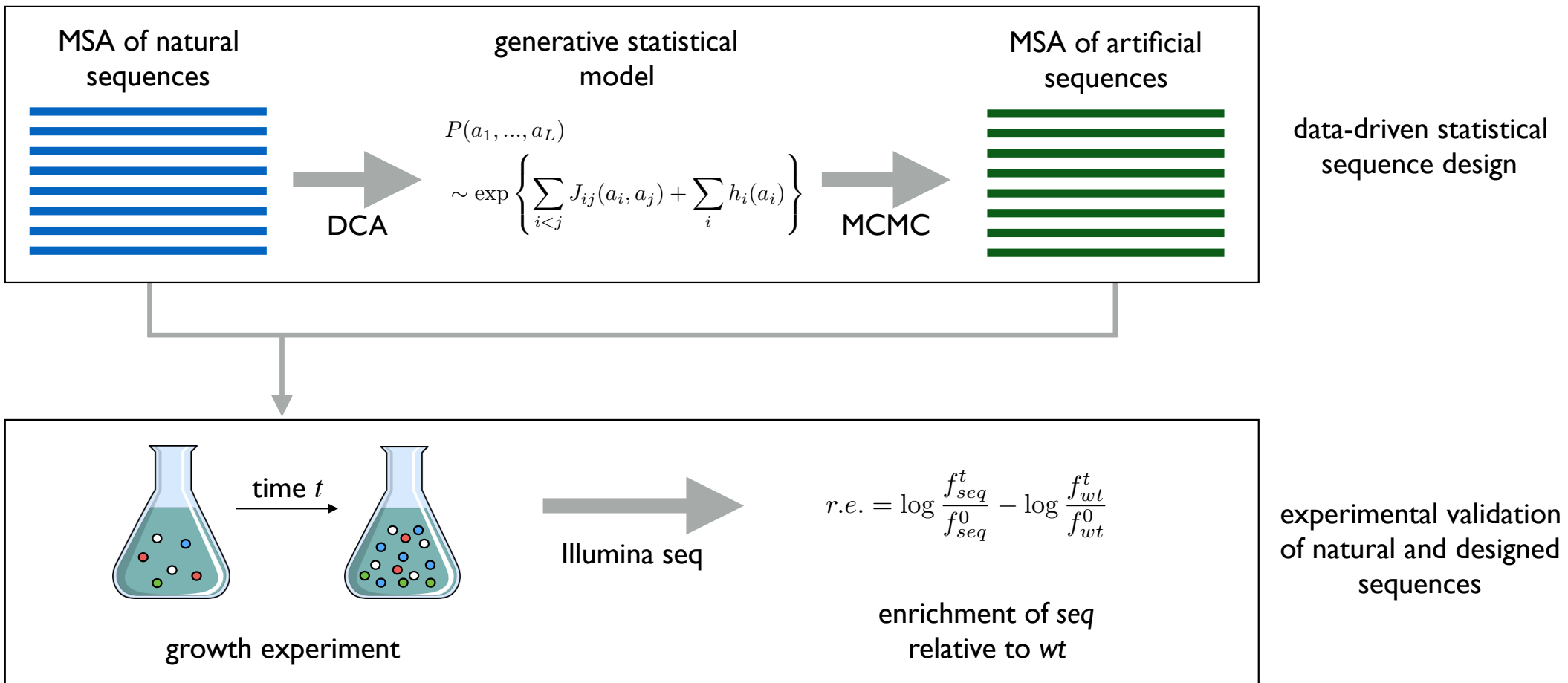
response regulator (PF00072)



➔ tested for many protein families

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

Coevolution-guided protein design



...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

Funding:

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)