# Randomness and information in biological data
# BIO-369

Prof. Anne-Florence Bitbol

EPFL

Lecture 12

# Outline of the course

**II Extracting information from biological data**

1 Quantifying randomness and information in data: entropy
- 1.1 Notion of entropy
- 1.2 Interpretation of entropy
- 1.3 Entropy in neuroscience data: response of a neuron to a sensory input

2 Quantifying statistical dependence
- 2.1 Covariance and correlation
- 2.2 Mutual information
- 2.3 Identifying coevolving sites in interacting proteins using sequence data

3 Inferring probability distributions from data
- 3.1 Model selection and parameter estimation: maximum likelihood
- 3.2 Introduction to maximum entropy inference
- 3.3 Predicting protein structure from sequence data

4 Finding relevant dimensions in data: dimension reduction
- 4.1 Principal component analysis
- 4.2 Beyond principal component analysis

Here we maximized entropy at fixed average energy. What do you think this procedure is equivalent to?

0%
0%
0%
0%

A. Maximizing the energy
B. Minimizing the energy
C. Maximizing the free energy
D. Minimizing the free energy

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Which of the following assertions is true?

0%

0%

0%

0%

A. $P(x) = \Sigma_{x,y} P(x,y)$

B. $P(x) = \Sigma_{y} P(x,y)$

C. $P(x) = \Sigma_{y} P(x|y)$

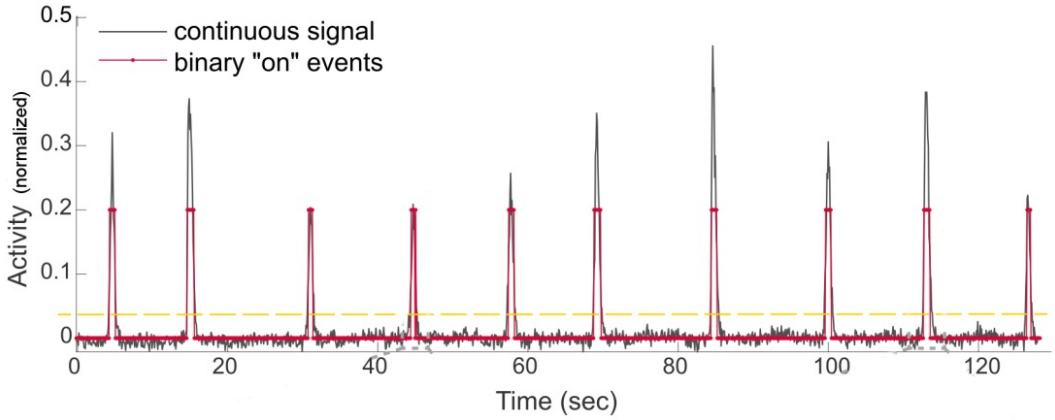
D. $P(x) = \Sigma_{y} P(y|x)$

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

# Some applications of maximum entropy modeling

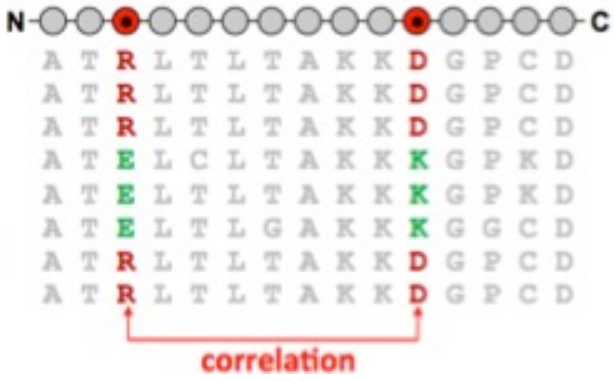
- **Neuroscience data:**




$$P(\{\sigma_i\}) = \frac{1}{Z}\exp[-E(\{\sigma_i\})].$$

$$E(\{\sigma_i\}) = -\sum_{i=1}^{N} h_i\sigma_i - \frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j$$

- **Protein sequence data:**




$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z}\exp\left\{-\left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j)\right]\right\}$$

# Outline of the course

**II Extracting information from biological data**

1 Quantifying randomness and information in data: entropy
- 1.1 Notion of entropy
- 1.2 Interpretation of entropy
- 1.3 Entropy in neuroscience data: response of a neuron to a sensory input

2 Quantifying statistical dependence
- 2.1 Covariance and correlation
- 2.2 Mutual information
- 2.3 Identifying coevolving sites in interacting proteins using sequence data

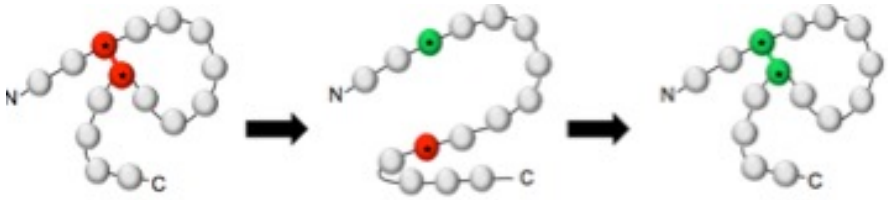
3 Inferring probability distributions from data
- 3.1 Model selection and parameter estimation: maximum likelihood
- 3.2 Introduction to maximum entropy inference
- 3.3 Predicting protein structure from sequence data

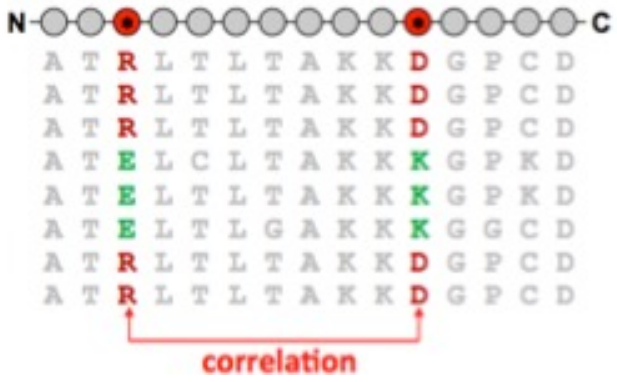
4 Finding relevant dimensions in data: dimension reduction
- 4.1 Principal component analysis
- 4.2 Beyond principal component analysis

# Protein sequence data

- **Inferring structure and function from sequences**

  - Data-driven approaches



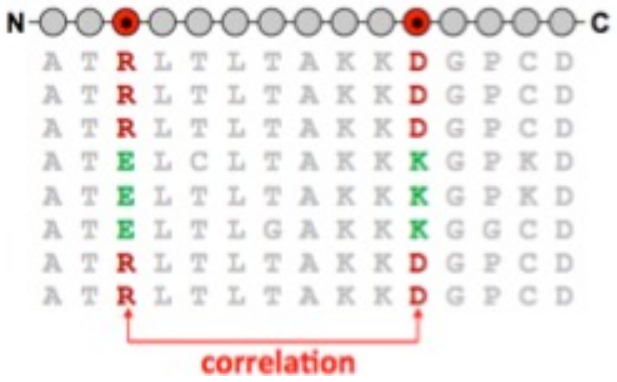

homologs - a protein family

Evolutionary coupling between interacting residues
→ correlations in multiple sequence alignments inform us about structure and function
BUT... observed correlations can be indirect A ↔ B ↔ C

# Maximum entropy model of protein sequence data

- **Goal: joint probability distribution**

  $P(\alpha_1, \alpha_2, \ldots, \alpha_L)$ probability of a sequence in the protein family




- **Observations retained: one- and two-body frequencies**

```
...ISHEL...
...VSHDI...
...VSHEL...
     ⋮
```

$$\rightarrow \begin{cases} f_i(\alpha) & i \in \{1, .., L\} \\ f_{ij}(\alpha, \beta) & \alpha \in \{A_1, .., A_{20}, A_{21} = -\} \end{cases}$$

- **Maximum entropy model consistent with these observations**

$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ - \left[ \sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\} \quad \rightarrow \text{Potts model}$$

one-body terms - fields

two-body terms - (direct) couplings

# Maximum entropy model of protein sequence data

- **Pairwise maximum entropy model and direct couplings:**

$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ - \left[ \sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\}$$

One needs to determine the fields and couplings consistent with the observations

$$\sum_{\boldsymbol{\alpha} \text{ such that } \alpha_i = \alpha} P(\boldsymbol{\alpha}) = f_i(\alpha)\,,$$

$$\sum_{\boldsymbol{\alpha} \text{ such that } \alpha_i = \alpha \text{ and } \alpha_j = \beta} P(\boldsymbol{\alpha}) = f_{ij}(\alpha, \beta)$$

→ very hard problem! (inverse problem)
→ many approximation methods
Cocco et al. (2017) - in the context of proteins

**Mean-field approximation:**

$$e_{ij}(\alpha, \beta) = C_{ij}^{-1}(\alpha, \beta)$$
$$C_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha) f_j(\beta)$$

(20 *L* x 20 *L* matrix)

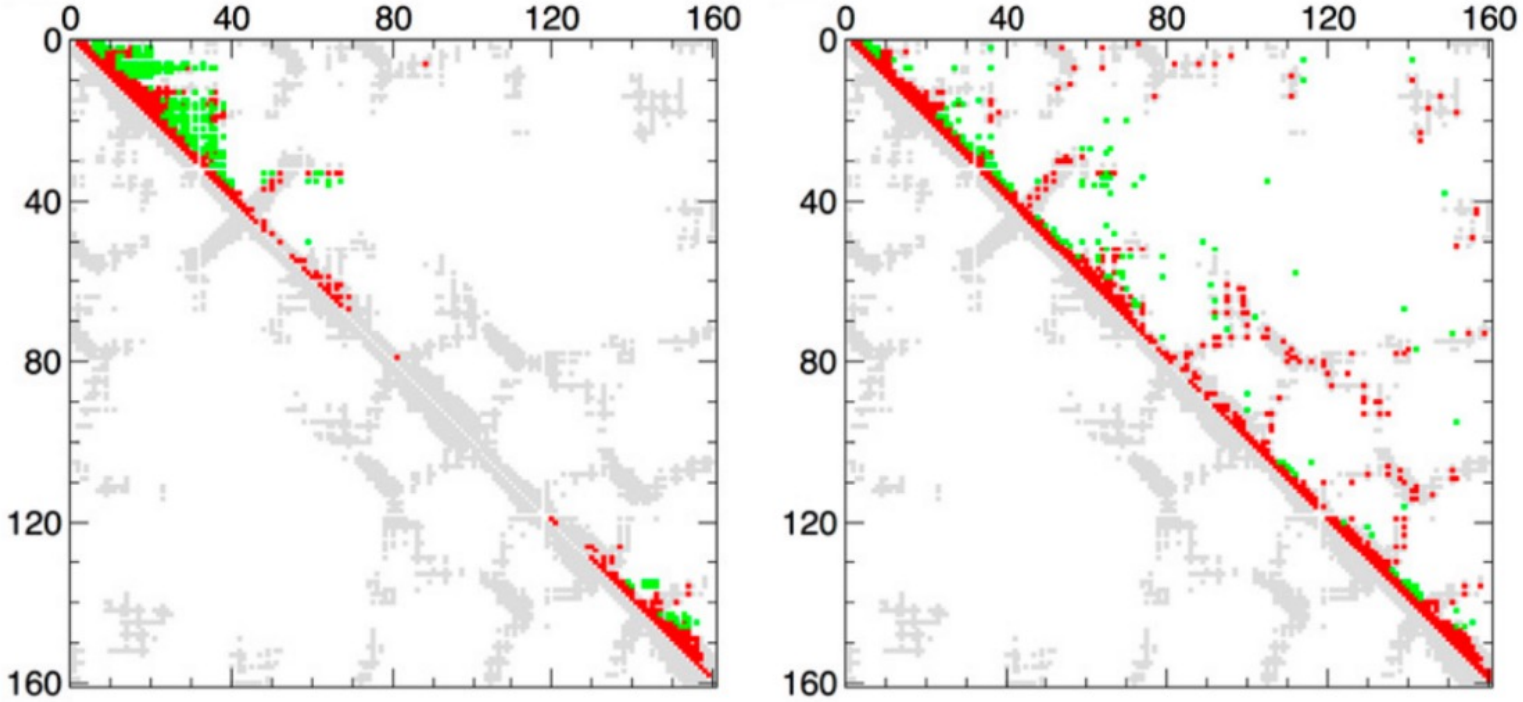
- Simplest approximation, can be derived through a small-coupling expansion
- Has proved rather good in the case of proteins

Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)

# Structure prediction

$e_{ij}(\alpha, \beta)$ much better predictor of 3D contact than | $C_{ij}(\alpha, \beta)$ Mutual Information

Weigt, White et al. (2009)
Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)




Contact map prediction for the eukaryotic signaling protein Ras

Mutual information (left)
Direct couplings (right)

Morcos, Pagnani et al (2011)

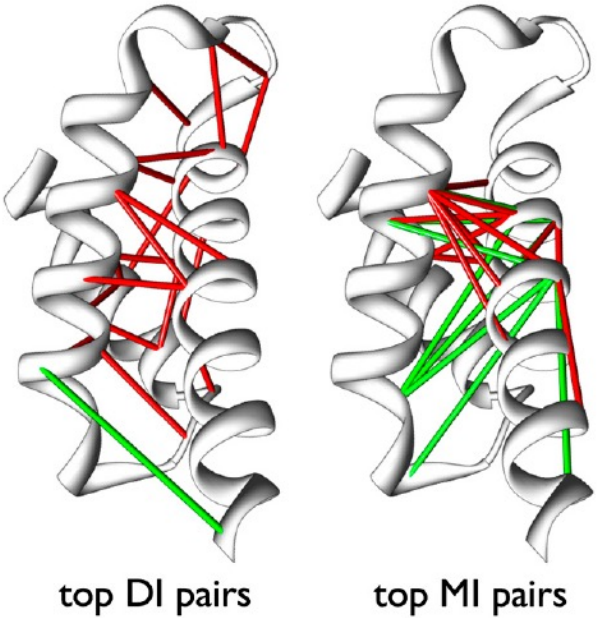
Gray: experimental contacts (cutoff: 8 A)
Red: correct predictions; green: incorrect ones

# Structure prediction

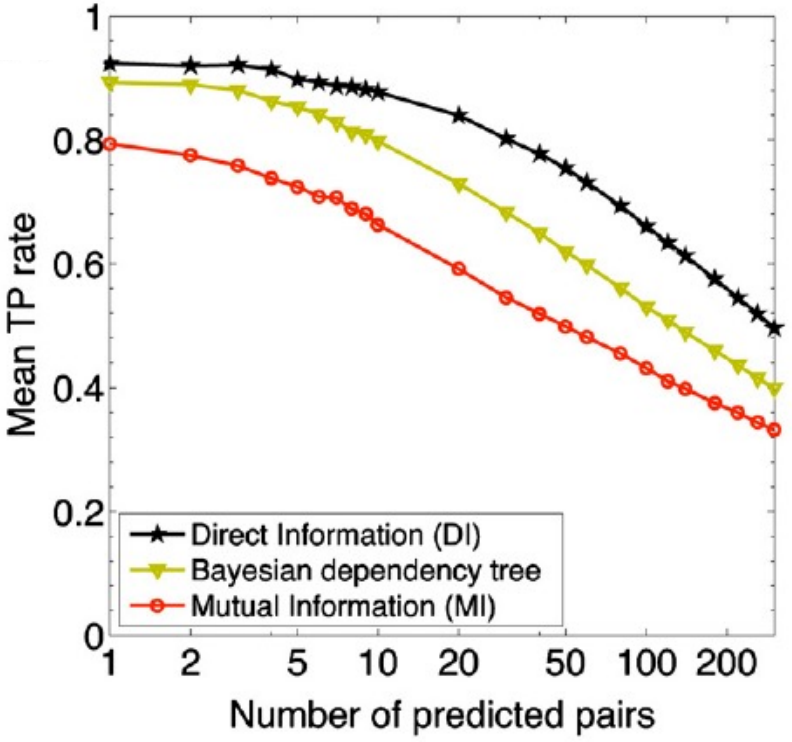
$e_{ij}(\alpha, \beta)$ much better predictor of 3D contact than $C_{ij}(\alpha, \beta)$

Mutual Information

Weigt, White et al. (2009)
Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)




Bacterial Sigma factor region 2.
Top 20 DI / MI predictions
(distance along the backbone > 4).
Red: distance <8 Å; green: others.

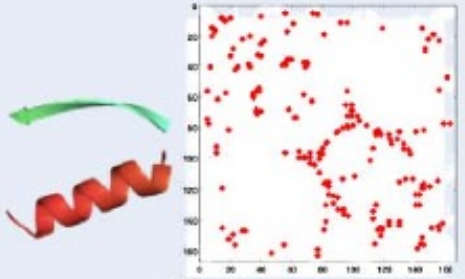



Mean TP rate for 131 domain families
vs. number of top-ranked contacts

# Structure prediction

Analyze the highest scoring pairs to produce ranked list of residue pairs which we predict to be close in 3D space. Use these pairs as predicted close "evolutionary inferred contacts" , EICs, in folding calculations

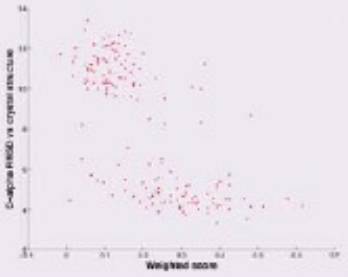
```
assign (resid 143 and name CA) (resid 123 and name CA) 4 4 3
assign (resid 16 and name CA) (resid 10 and name CA) 4 4 3
assign (resid 141 and name CA) (resid 82 and name CA) 4 4 3
assign (resid 129 and name CA) (resid 87 and name CA) 4 4 3
assign (resid 92 and name CA) (resid 11 and name CA) 4 4 3
assign (resid 116 and name CA) (resid 81 and name CA) 4 4 3
```



predicted contacts (EICs)

Start with extended structure
use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain

Rank predicted structures using quality measure of backbone alpha torsion and beta sheet twist

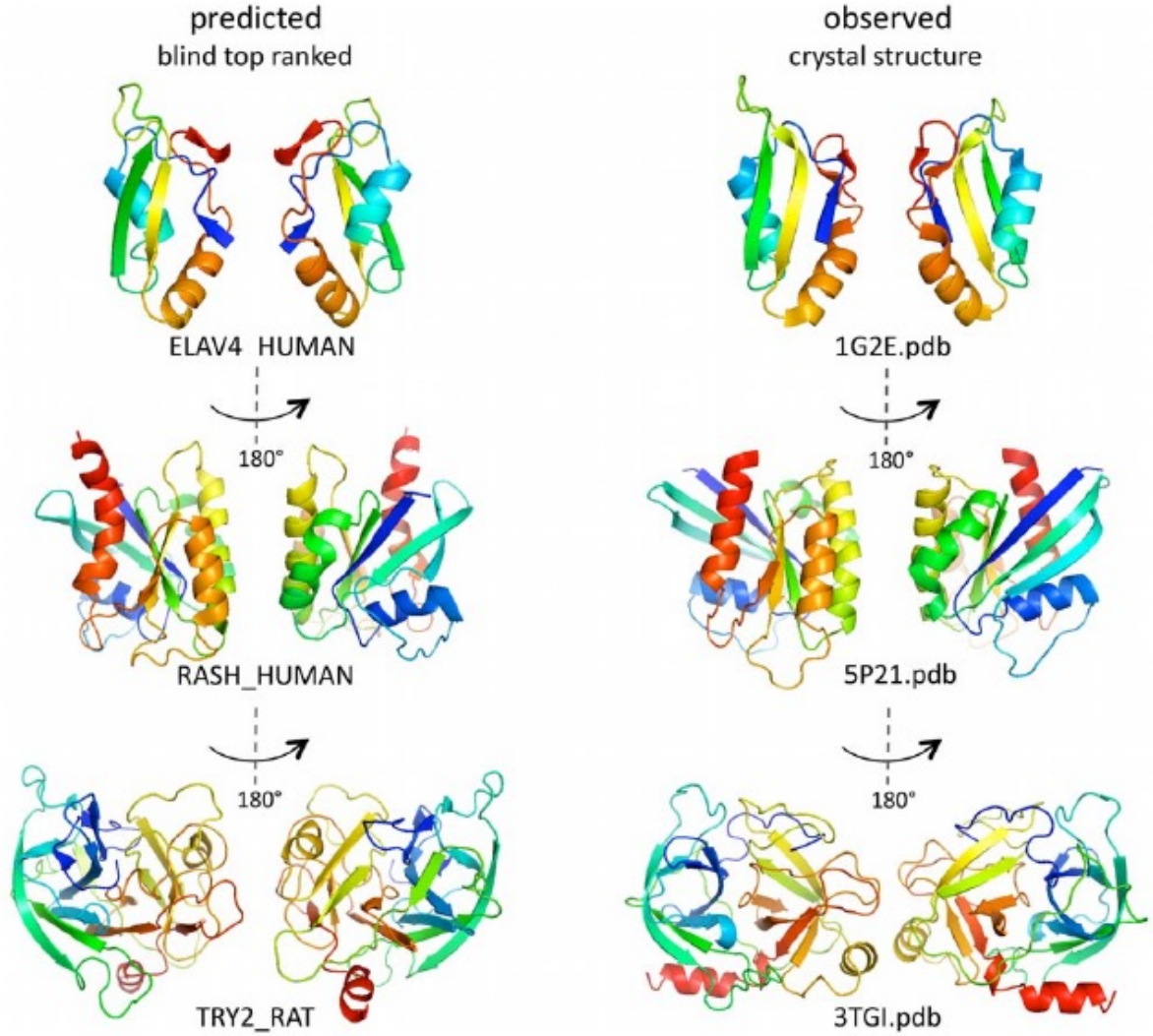


good scores

bad scores

Marks, Colwell et al (2011)

# Structure prediction




Results for 3 proteins:
- predicted top ranked 3D structure (left)
- experimentally observed structure (right)

Each structure in front and back view

Marks, Colwell et al (2011)

# Maximum entropy model of sequence data

- **Limitations of the structure prediction method**
  - Requires large alignments of homologous proteins (~ a few hundreds)
  - Requires a high diversity within these alignments

  → cannot be used for small protein families (recall: one model *per family*)

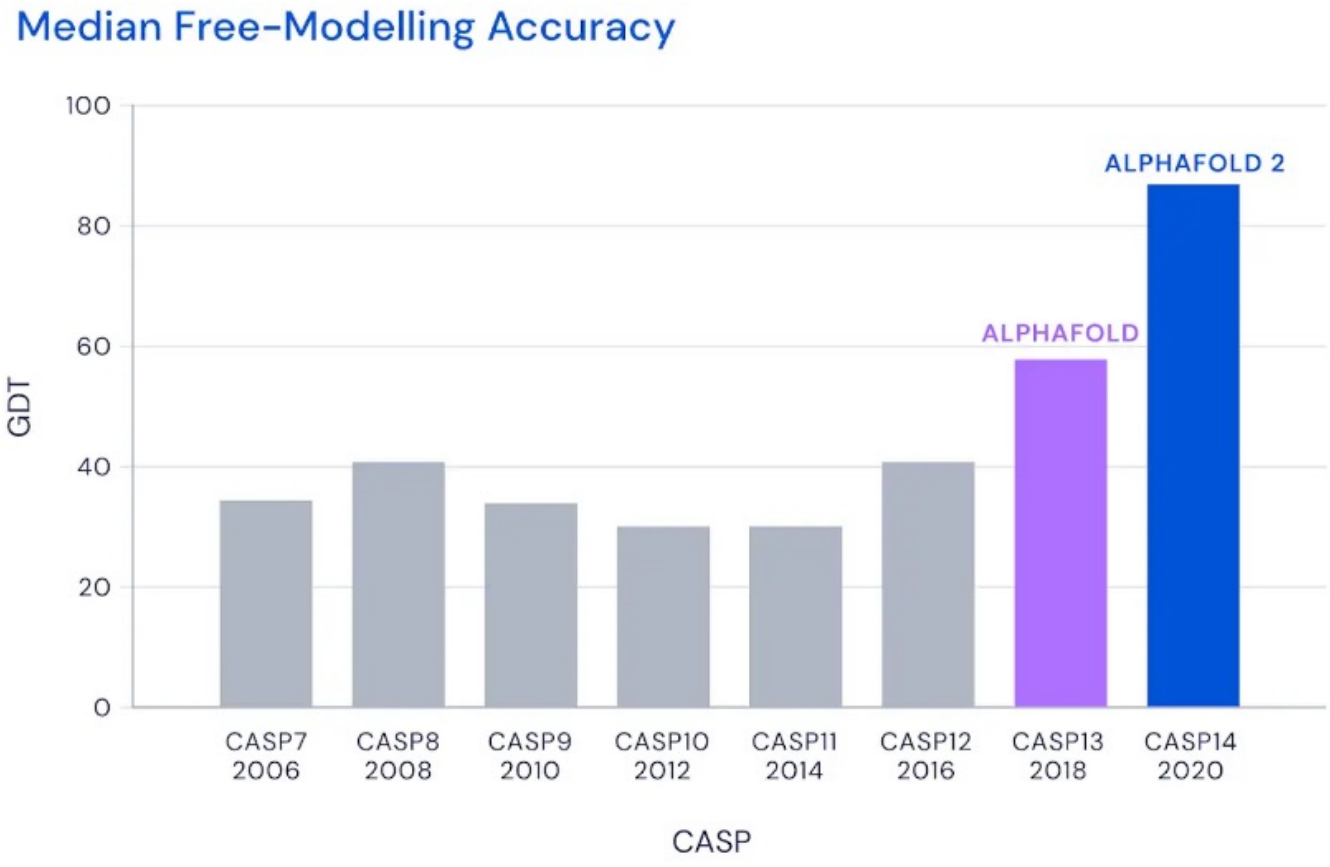
- **Other applications of the maximum entropy model for protein sequences**
  - Mutation effect prediction
  - Protein-protein interaction prediction
  - Protein design

- **Conclusion on the applications of maximum entropy models**
  - Neuroscience data (infer connections, predict pattern rates, study collective behavior)
  - Protein sequences (predict structure, mutation effects, interactions, model evolution)

  But also:
  - Reconstruction of signaling pathways from expression data
  - Predicting response to multidrug combinations in bacteria, etc.
  - Useful in signal processing (ex. MNR data)

# Recent developments in protein structure prediction

- **CASP14 and AlphaFold2**




GDT: global distance test

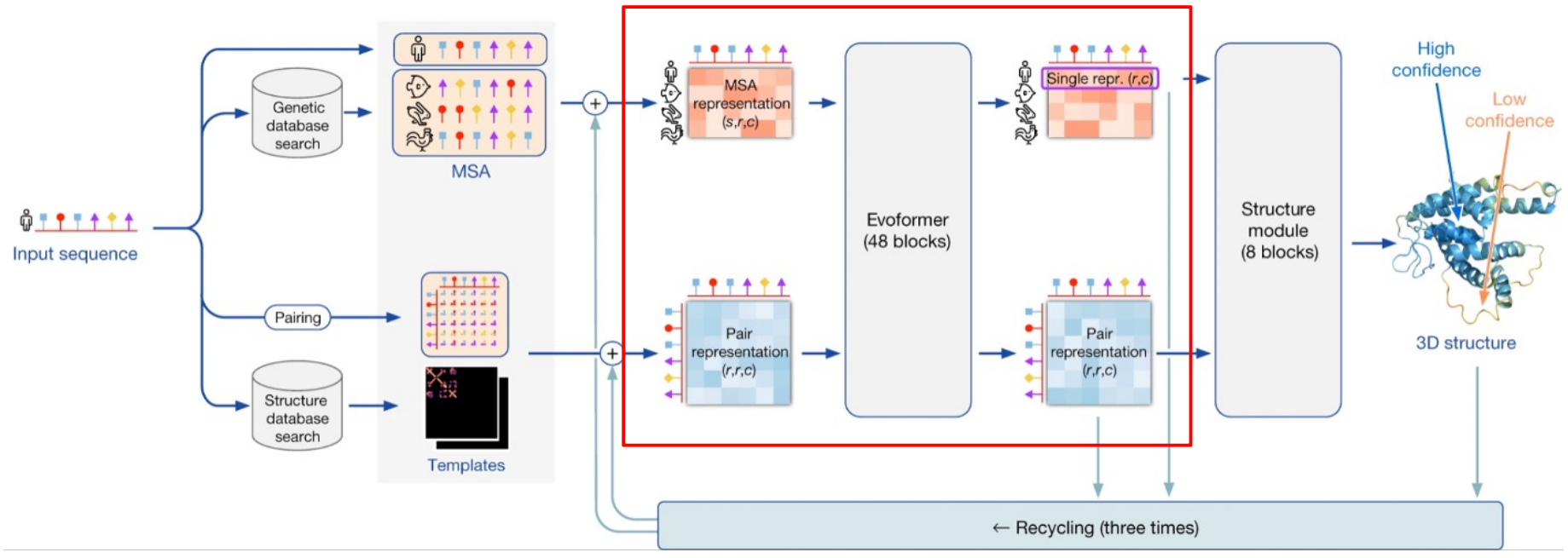
median score: 92.4 GDT
→ RMSD ~ 1.6 A

Free-modelling category (hardest): median score 87.0 GDT

# A few words about AlphaFold

## Architecture of AlphaFold2 – Jumper et al 2021

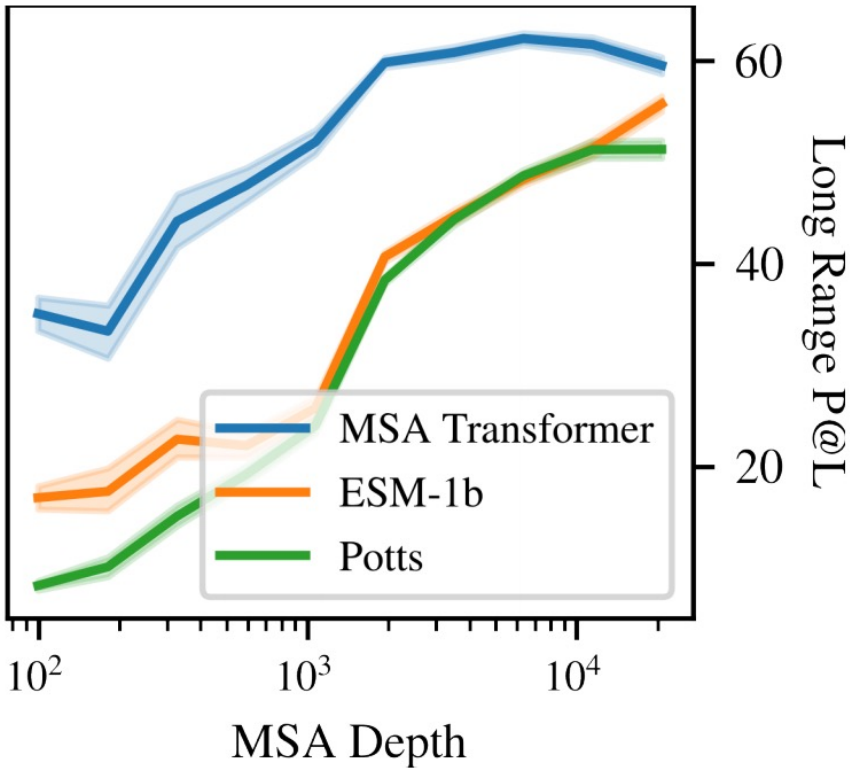
- Deep learning approach – one large model for many protein families
- Starts from multiple sequence alignments of homologous sequences & from structures
- Uses natural language processing methods:
  - Attention (Bahdanau et al 2014), transformer architecture (Vaswani et al 2017)
  - Specifically, part of AlphaFold is a **protein language model trained on MSAs**




Jumper et al 2021

# Recent developments in protein structure prediction

- **Recent unsupervised models (transformers)**



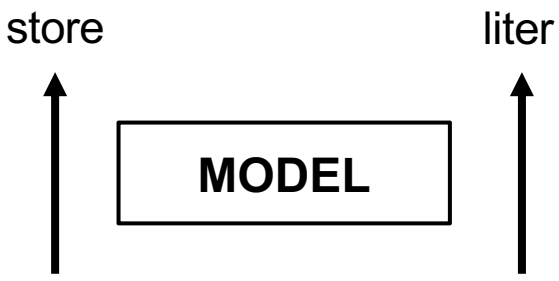

- One model *for all families / all proteins*
- Performs better than maximum entropy models for alignments with relatively few sequences

Rives et al 2021

# Sequence-based module (EvoFormer): inspired by natural language processing

- **Masked Language Modeling objective: self-supervised learning –** Devlin et al 2018

Randomly **mask** a fraction of the **words** and train the model to predict them using the surrounding **context**




The man went to the **[MASK]** and bought a **[MASK]** of milk.

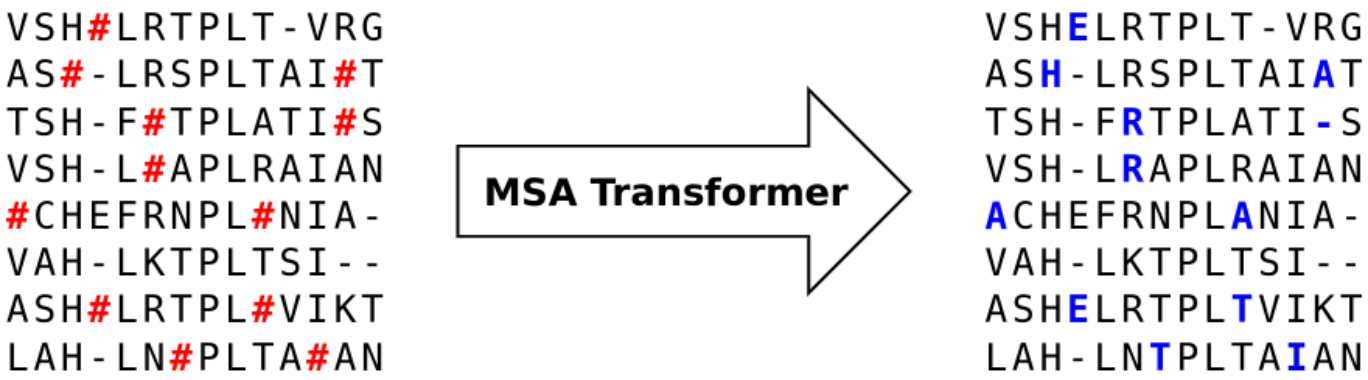
The model is trained to minimize a pseudo-likelihood loss:

$$L_{MLM}(x,\theta) = -\sum_{m \in mask} \log p(x_m \mid \tilde{x};\theta) \qquad with \;\; \tilde{x}: masked\ sentence$$

# MSA Transformer (similar to AlphaFold's EvoFormer, but not supervised)

- **Masked Language Modeling (MLM) objective on protein MSAs –** Rao et al 2021

Randomly mask (#) a fraction of the amino acids and train the model to predict them, using the surrounding context




The model is trained to minimize a pseudo-likelihood loss:

$$\mathcal{L}_{\mathrm{MLM}}(\mathcal{M}, \widetilde{\mathcal{M}}; \theta) = -\sum_{(m,i) \in \mathrm{mask}} \log p(x_{m,i} \mid \widetilde{\mathcal{M}}; \theta)$$

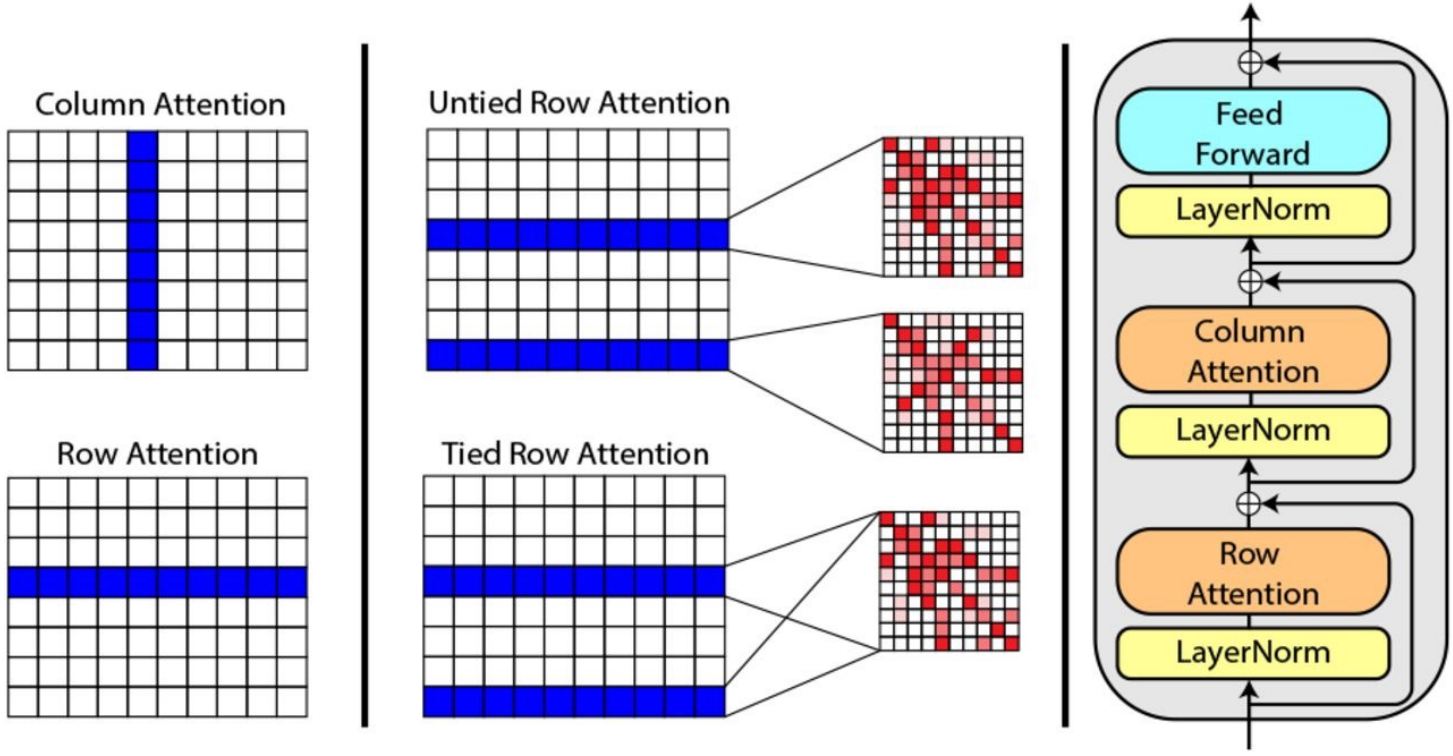
$\mathcal{M}$ MSA

$\widetilde{\mathcal{M}}$ masked MSA

MSA Transformer is similar to AlphaFold's EvoFormer, but it is self-supervised

# Architecture of MSA Transformer

- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021

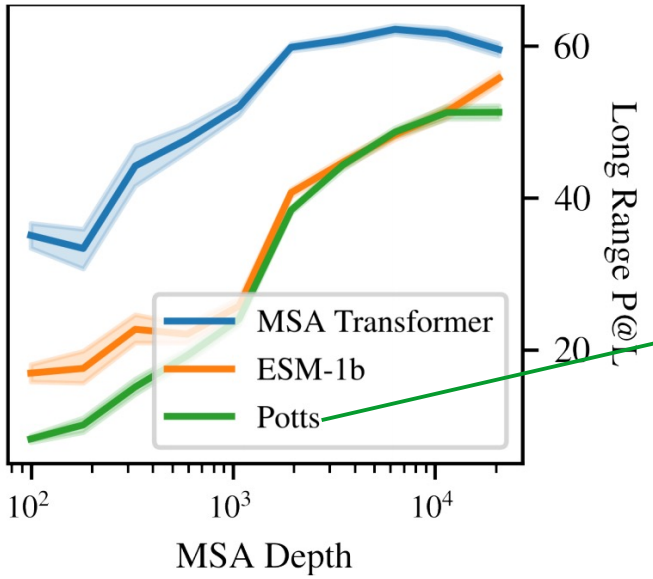



Training set:
- 26M MSAs corresponding to UniRef50 clusters
- average depth of MSAs: 1192

# Unsupervised structural contact prediction by MSA Transformer

- **(Tied) row attentions capture structural contacts –** Rao et al 2021

  - Simple combinations of the row attention softmax matrices allow contact prediction
  - State-of-the-art unsupervised contact prediction




Contact prediction performance

Potts model: pairwise maximum entropy model / DCA [Weigt, White et al 2009]

$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ -\left[ \sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\}$$

One model per family
(vs. language models trained on many families)

For unsupervised contact prediction, MSA Transformer outperforms:
- Potts models
- BERT-like single-sequence models (ESM-1b, still true with ESM-2)

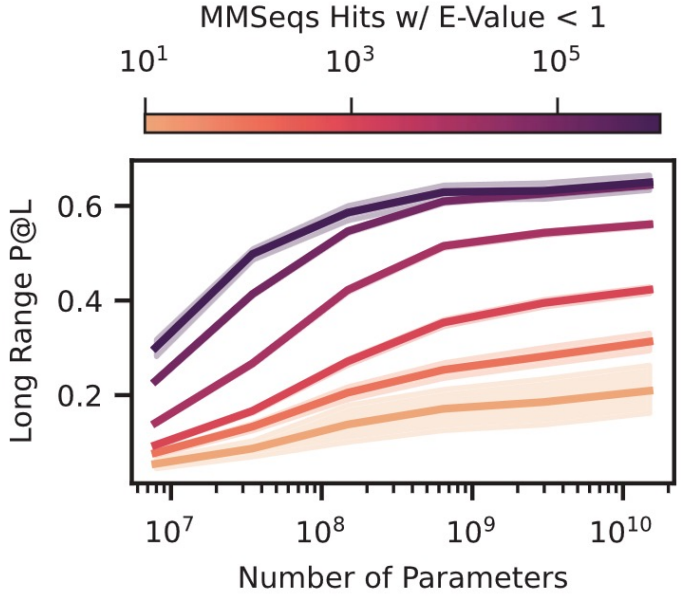
# Are MSAs really necessary?

## Structure prediction based on single-sequence language models

**Motivations:**
- Some proteins have few homologs
- MSA construction is imperfect and slow
- Predicting structure from a single sequence would bring us closer to “understanding protein folding”

**Strategy:**
- Train language models on large ensembles of non-aligned single sequences
- Add a structure module inspired by the one of AlphaFold2

AminoBERT → RGN2 (Chowdhury et al 2021); OmegaPLM → OmegaFold (Wu et al 2022);
ESM-2 → ESMFold (Lin et al 2023)




ESM-2 & ESMFold (Lin et al 2023):

**(Unsupervised) contact prediction:**
- slightly less good than with MSA Transformer, even with many more parameters (15B vs. 100M)
- still very strongly affected by the number of existing homologs!

**(Supervised) structure prediction:**
- less good than AlphaFold2
- much faster – enabled structure prediction at metagenomic scale