Making sense of protein sequence data using machine learning

Anne-Florence Bitbol

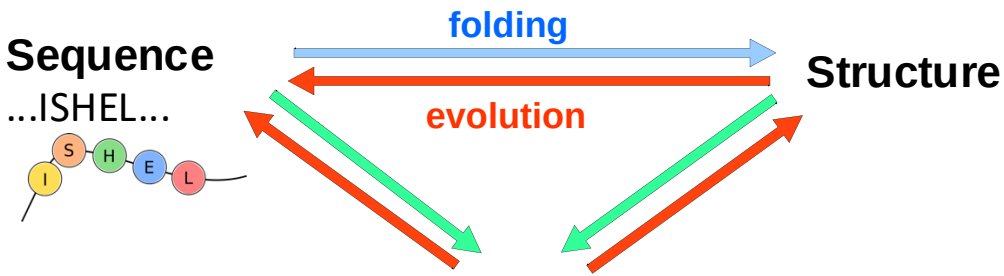
EPFL

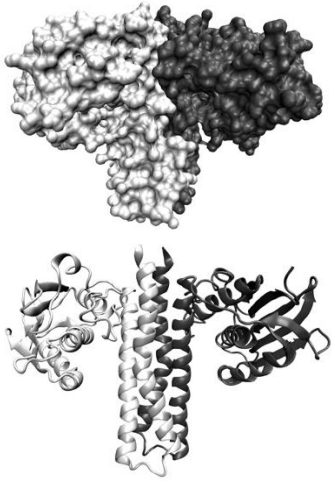
PHYS-754
September 12, 2024

Part 1
Inference from protein sequences: traditional methods

Introduction

Proteins





Mutations act on sequences
BUT
selection acts on function

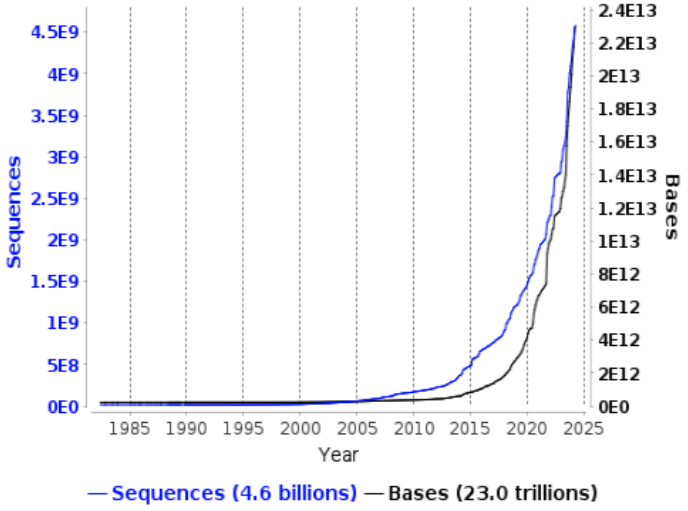
- Heteropolymers made of 20 types of amino-acids (monomers) → ~20^{100} possible proteins
- A given natural protein folds into a compact and (almost) unique 3D **structure**
- It has specific **interactions** with other molecules → **function**

- Experiment: random proteins do not fold properly Socolich et al. (2005)

→ Natural proteins are special, due to natural selection for folding and function

Introduction

- **A growing amount of sequence data**



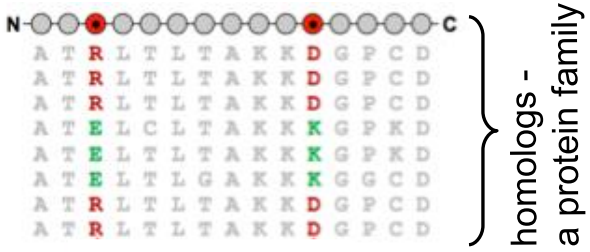
Accumulating sequence data
(currently > 10^9 sequences)

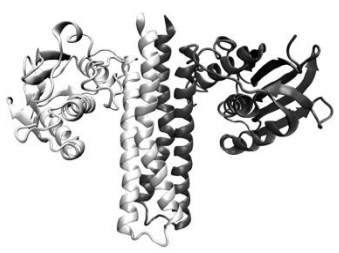
→ Great opportunity for statistical physics, information theory and machine learning methods to learn about proteins!

Goals: infer structure, function, interactions

https://www.ebi.ac.uk/ena/browser/about/statistics

- **Protein families and multiple sequence alignments (MSAs)**

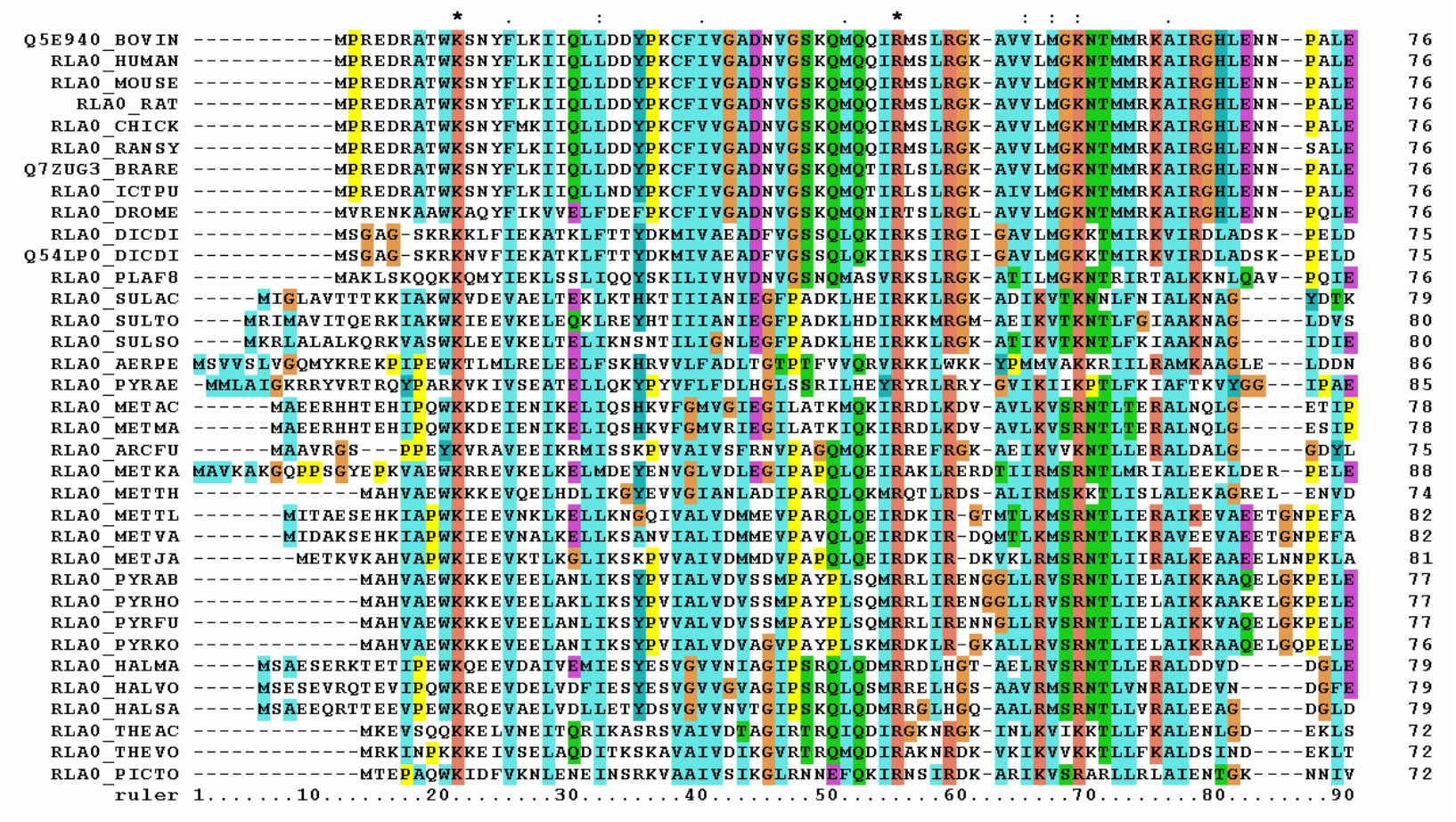




Introduction – reminder

- **Multiple sequence alignments (MSAs)**

Focus on amino-acid sequences of proteins (translated from the coding part of genomes)



Acidic ribosomal protein P0 (first 90 positions) from several organisms

Row = sequence
Column = site (given position in 3D structure)

Colors = level of conservation

Conservation in MSAs

- **Some columns are highly conserved**

```
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-QKQFVSDASHELRTPISVIQGYIDLLDRDKEVLEEAIEAIQAETTSMKKLLEQLLFLARSDKG
-RKELIANISHDLKTPITAIKGYVEGIRDSPEKLSRYVDTIYRKILEVDGLIDELFLFSKLD--
-KSEIIAMVSHELKTPLTSILAFGEILLALLPWQKEYLEDIMESGQELLKQIETLLTMAKIEAG
-----LHSLVHDLKTPLMTIQGLSSLIGLDSPKLQEYVQKIEQAVENVNKMISEIL--------
-RREFLANVSHELRTPLTIIQGYTEALLDTDEKIREHLKNILQEAERLKAMANELLDLASIEEG
-LGLLAAGVAHEINNPLATVSAYAEDLLERSGELARYLQVIGKQIERCKKITGSLLNFARQPA-
MRSEFIANVSHELRTPLTSIKGFLETLLDDKTIAKHFLQIMNSETERLTRLIDDLLSLSKIEA-
-RRQMIADIAHELRTPLSILQGNFELLLEVIEADEETLRSLAEEVKRLSRLVEELRELSLAEAG
-QKEFFANVSHELRSPATAILGEAQITLRSDDEYRQTLLRISESAEQLAFRIEDLLMLIRHDE-
```

Conservation in MSAs

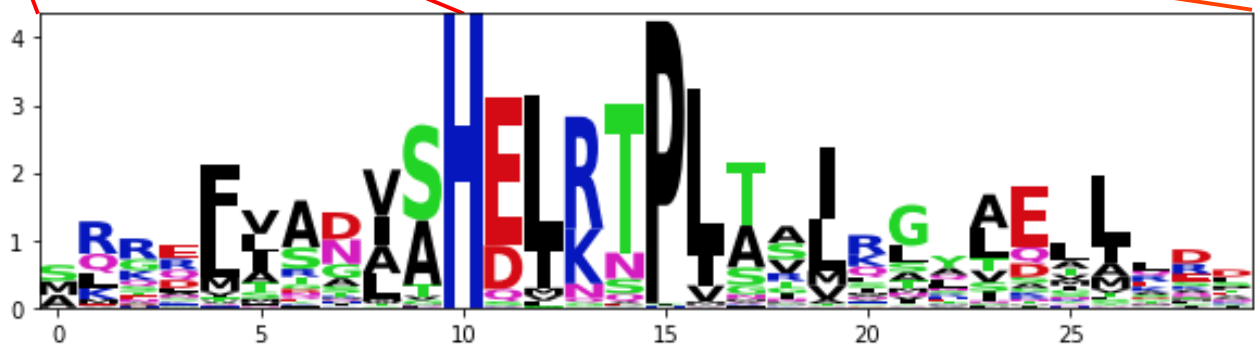
- **Some columns are highly conserved**

```
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-QKQFVSDASHELRTPISVIQGYIDLLDRDKEVLEEAIEAIQAETTSMKKLLEQLLFLARSDKG
-RKELIANISHDLKTPITAIKGYVEGIRDSPEKLSRYVDTIYRKILEVDGLIDELFLFSKLD--
-KSEIIAMVSHELKTPLTSILAFGEILLALLPWQKEYLEDIMESGQELLKQIETLLTMAKIEAG
-----LHSLVHDLKTPLMTIQGLSSLIGLDSPKLQEYVQKIEQAVENVNKMISEIL--------
-RREFLANVSHELRTPLTIIQGYTEALLDTDEKIREHLKNILQEAERLKAMANELLDLASIEEG
-LGLLAAGVAHEINNPLATVSAYAEDLLERSGELARYLQVIGKQIERCKKITGSLLNFARQPA-
MRSEFIANVSHELRTPLTSIKGFLETLLDDKTIAKHFLQIMNSETERLTRLIDDLLSLSKIEA-
-RRQMIADIAHELRTPLSILQGNFELLLEVIEADEETLRSLAEEVKRLSRLVEELRELSLAEAG
-QKEFFANVSHELRSPATAILGEAQITLRSDDEYRQTLLRISESAEQLAFRIEDLLMLIRHDE-
```

Conservation in MSAs

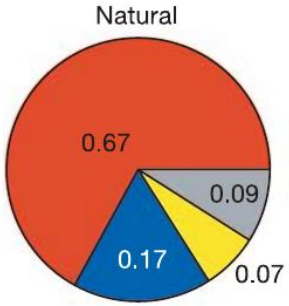
- **Some columns are highly conserved**

```
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-QKQFVSDASHELRTPISVIQGYIDLLDRDKEVLEEAIEAIQAETTSMKKLLEQLLFLARSDKG
-RKELIANISHDLKTPITAIKGYVEGIRDSPEKLSRYVDTIYRKILEVDGLIDELFLFSKLD--
-KSEIIAMVSHELKTPLTSILAFGEILLALLPWQKEYLEDIMESGQELLKQIETLLTMAKIEAG
-----LHSLVHDLKTPLMTIQGLSSLIGLDSPKLQEYVQKIEQAVENVNKMISEIL--------
-RREFLANVSHELRTPLTIIQGYTEALLDTDEKIREHLKNILQEAERLKAMANELLDLASIEEG
-LGLLAAGVAHEINNPLATVSAYAEDLLERSGELARYLQVIGKQIERCKKITGSLLNFARQPA-
MRSEFIANVSHELRTPLTSIKGFLETLLDDKTIAKHFLQIMNSETERLTRLIDDLLSLSKIEA-
-RRQMIADIAHELRTPLSILQGNFELLLEVIEADEETLRSLAEEVKRLSRLVEELRELSLAEAG
-QKEFFANVSHELRSPATAILGEAQITLRSDDEYRQTLLRISESAEQLAFRIEDLLMLIRHDE-
```



Correlations in MSAs

Correlations in amino acid usage are crucial



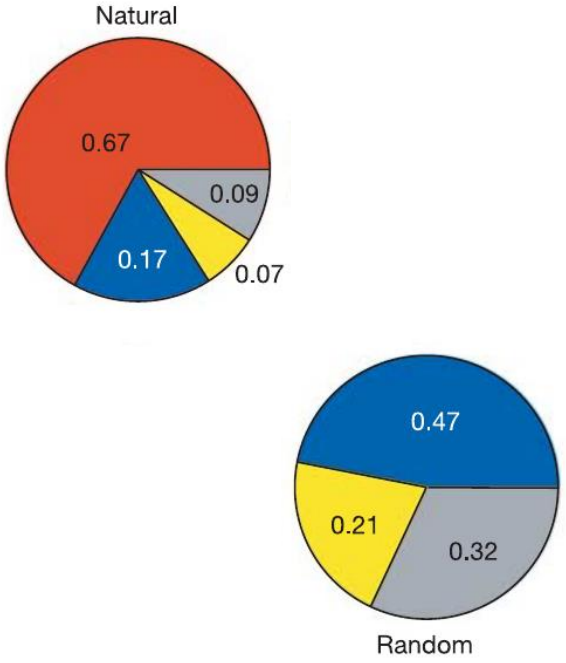
Socolich et al. 2005
synthetic WW domain

- Red, natively folded
- Blue, soluble but unfolded
- Yellow, insoluble
- Gray, poorly expressing

- Most natural sequences fold

Correlations in MSAs

- **Correlations in amino acid usage are crucial**



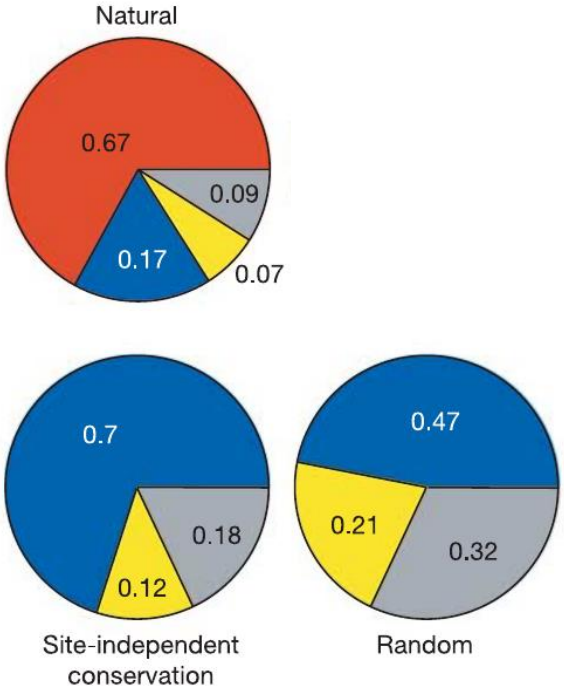
- Most natural sequences fold
- Random sequences don't

Socolich et al. 2005
synthetic WW domain

- Red, natively folded
- Blue, soluble but unfolded
- Yellow, insoluble
- Gray, poorly expressing

Correlations in MSAs

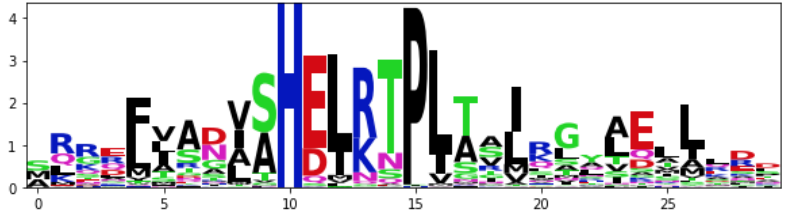
Correlations in amino acid usage are crucial



Socolich et al. 2005
synthetic WW domain

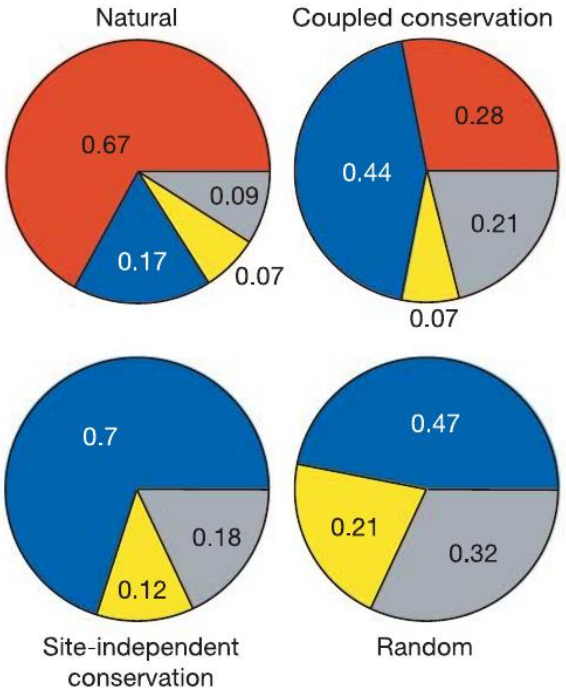
- Red, natively folded
- Blue, soluble but unfolded
- Yellow, insoluble
- Gray, poorly expressing

- Most natural sequences fold
- Random sequences don't
- Sequences reproducing the natural one-site frequencies don't



Correlations in MSAs

Correlations in amino acid usage are crucial



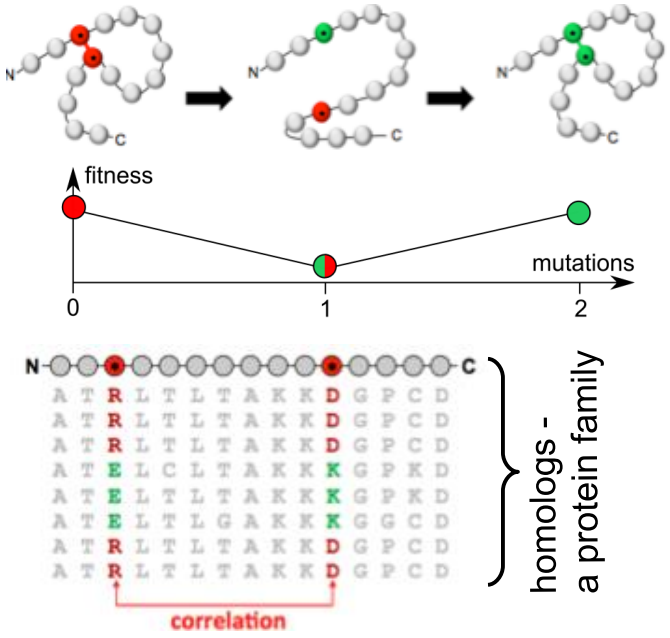
Socolich et al. 2005
synthetic WW domain

- Red, natively folded
- Blue, soluble but unfolded
- Yellow, insoluble
- Gray, poorly expressing

- Most natural sequences fold
- Random sequences don't
- Sequences reproducing the natural one-site frequencies don't
- Some sequences reproducing conserved correlations in addition do!

Protein sequence data and inference

- **Inferring structure and function from sequences**



Evolutionary coupling between interacting residues
→ correlations in MSAs inform us about
structure and function

Several approaches exploit these signatures to understand protein structure, interactions and function

de Juan et al, 2013

- **Simple data-driven approach: retain some statistics**

One- and two-body frequencies; (generalized) covariances

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

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

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

Protein sequence data and inference

Information theory: quantifying conservation and statistical dependence

One- and two-body frequencies; (generalized) covariances

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

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

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

- **Shannon entropy:** quantifies conservation in an MSA column

$$H_i = -\sum_\alpha P_i(\alpha) \log\left[P_i(\alpha)\right] \approx -\sum_\alpha f_i(\alpha) \log\left[f_i(\alpha)\right]$$

  - 0 for fully conserved column
  - log(21) for uniformly chosen amino acid (or gap)

- **Mutual information:** quantifies statistical dependence between 2 MSA columns

$$MI_{ij} = \sum_{\alpha,\beta} P_{ij}(\alpha,\beta) \log\left[\frac{P_{ij}(\alpha,\beta)}{P_i(\alpha)P_j(\beta)}\right] \approx \sum_{\alpha,\beta} f_{ij}(\alpha,\beta) \log\left[\frac{f_{ij}(\alpha,\beta)}{f_i(\alpha)f_j(\beta)}\right]$$

  - Non-negative
  - 0 for statistically independent columns (and only then)

Limitations of covariance and mutual information

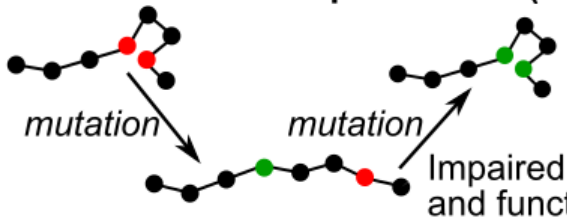
- **Issues**

Evolutionary coupling between interacting residues (coevolution)
→ pairwise correlations in multiple sequence alignments inform us about structure and function

But (1) observed correlations can be indirect A ↔ B ↔ C

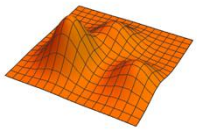
But (2) not all correlations come from functional constraints

Correlations from optimization (maintain structure and function):

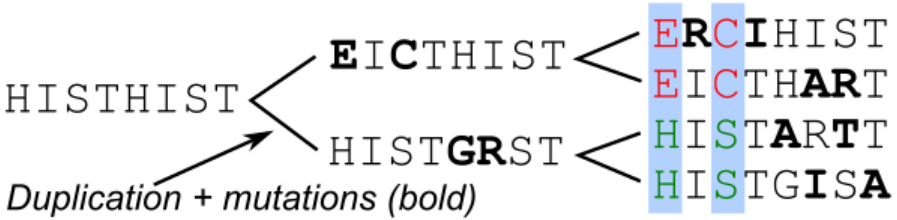


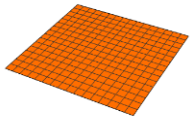
Observed sequences
(survived selection):

INTERACT
ANTGRANT
INAEPICT
INAGPANT



Correlations from historical contingency (phylogeny):





Potts models – Weigt, White et al, 2009

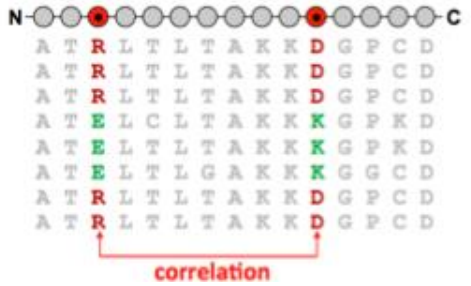
- **Goal: construct a global model for the protein family**

  Probability of observing a sequence in the family: $P(\alpha_1, \alpha_2, \ldots, \alpha_L)$

- **Construct it from the data (data-driven approach)**
  - Observations retained: one- and two-body frequencies (choice)

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

  - Multiple choices are consistent with these observations...



- **Maximum entropy principle**
  - Maximize $S = -\sum_{\{\alpha_1,\ldots,\alpha_L\}} P(\alpha_1, \ldots, \alpha_L) \log\left[P(\alpha_1, \ldots, \alpha_L)\right]$ under constraints
  - Yields the least-structured model consistent with the observations

- **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

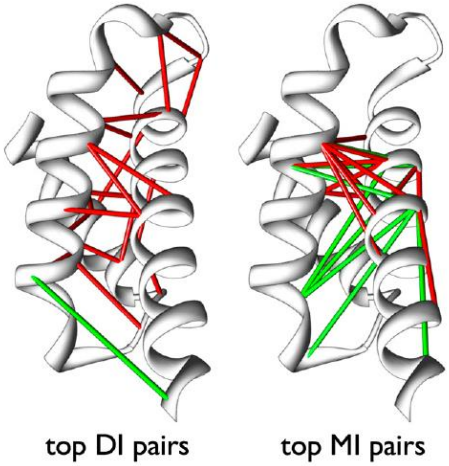
two-body terms - (direct) couplings

Structure prediction by Potts models

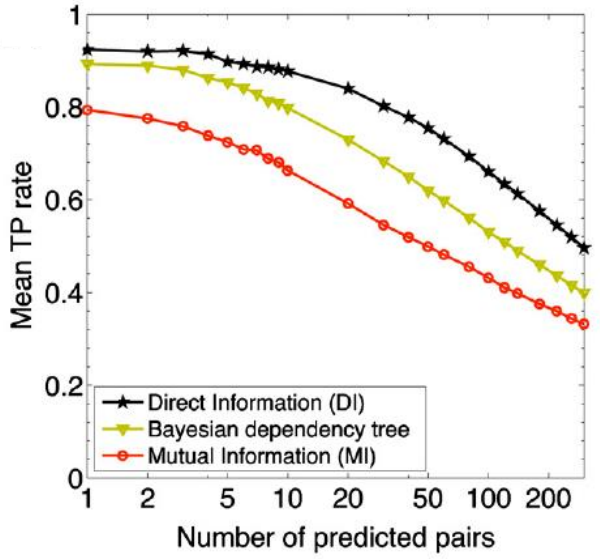
$e_{ij}(\alpha,\beta)$ much better predictor of 3D contact than $C_{ij}(\alpha,\beta)$ | Mutual Information

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

Morcos, Pagnani et al. (2011):



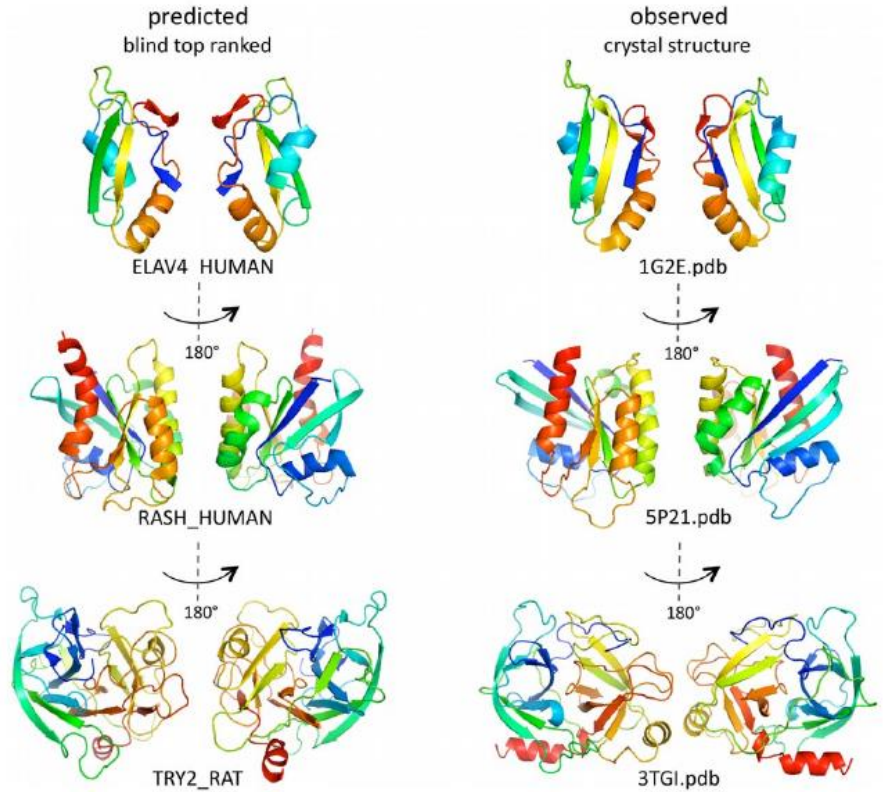
Bacterial Sigma factor region 2.
Top 20 DI / MI predictions
(distance along the backbone > 4).
Red: distance <8 Å; green: others.



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

Structure prediction by Potts models

Marks, Colwell et al. (2011):



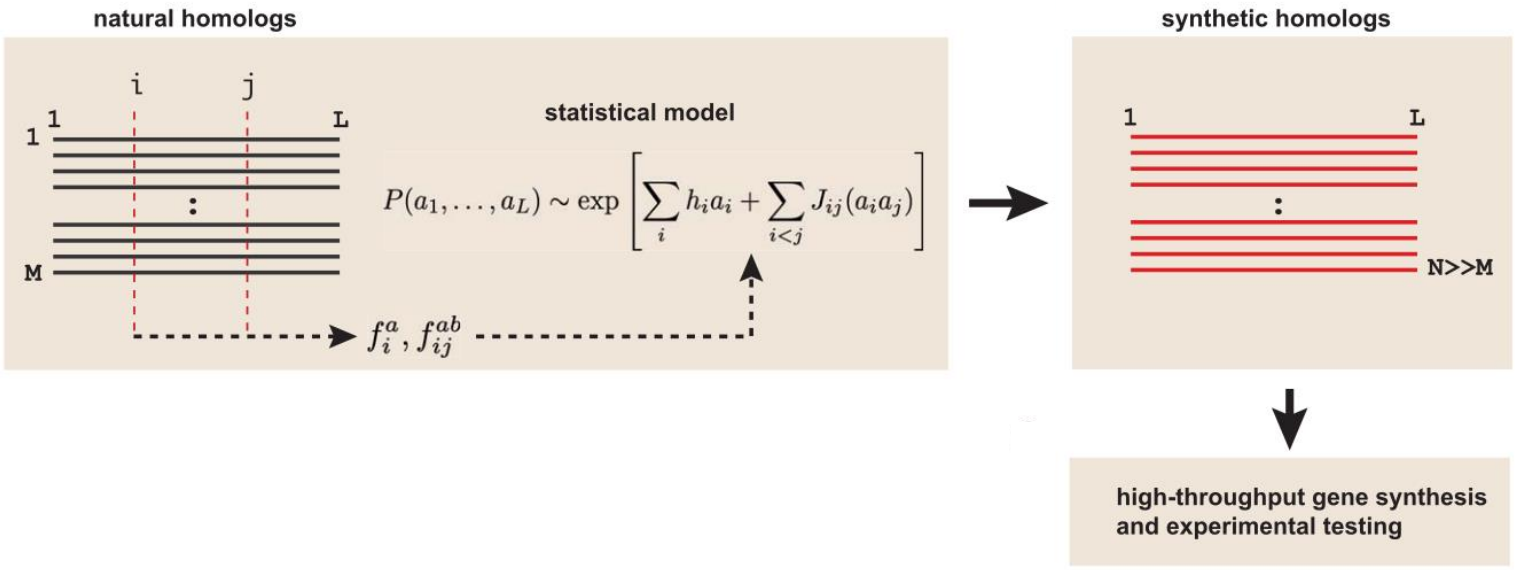
Results for 3 proteins:
- predicted top ranked 3D structure (left)
- experimentally observed structure (right)

Each structure in front and back view

Limitation: requires many diverse homologs

Beyond structure prediction: application to protein design

- **Potts models are generative – Russ et al 2020 – PAPER**
  - Using Potts models to generate new chorismate mutase enzymes

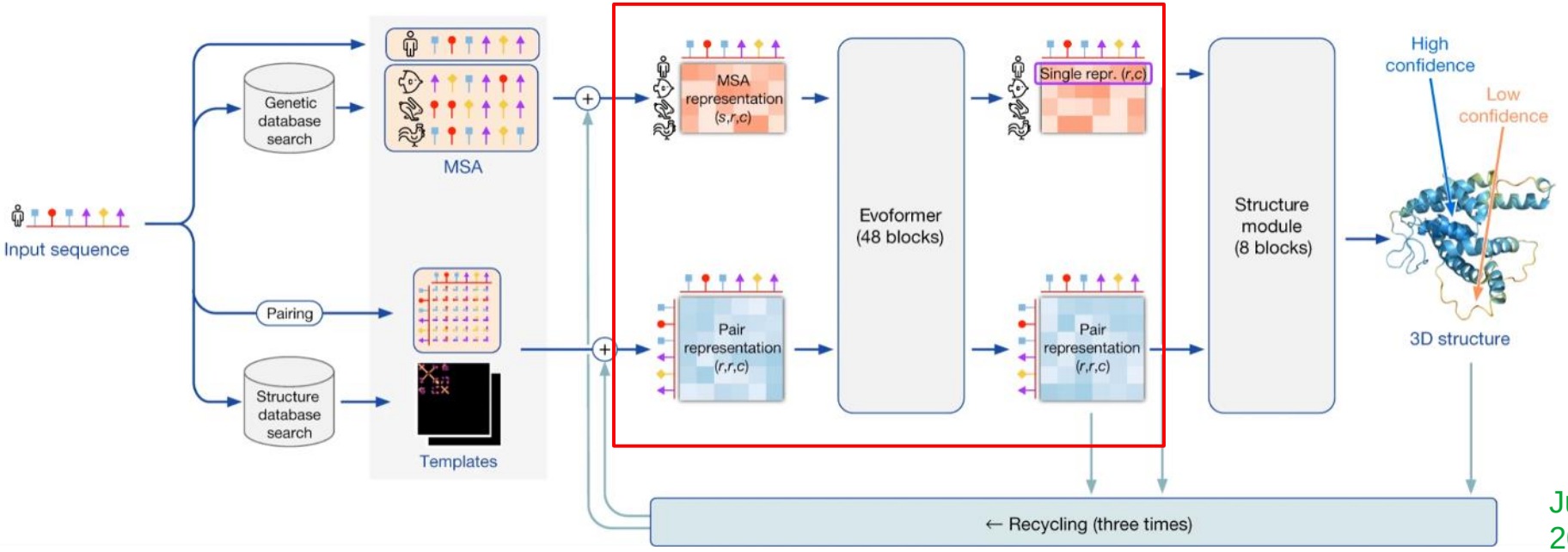


The model built on natural CM homologs is used to generate artificial sequences that were tested in a high-throughput assay for desired functions

Part 2
Inference from protein sequences: protein language models

A few words about AlphaFold

- **Recent developments in protein structure prediction – Jumper et al 2021**
  - **Supervised** deep learning approaches – AlphaFold, AlphaFold2 – won CASP13 and **CASP14**
    Other model: RoseTTAFold (Baek et al 2021)
  - AlphaFold2 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**

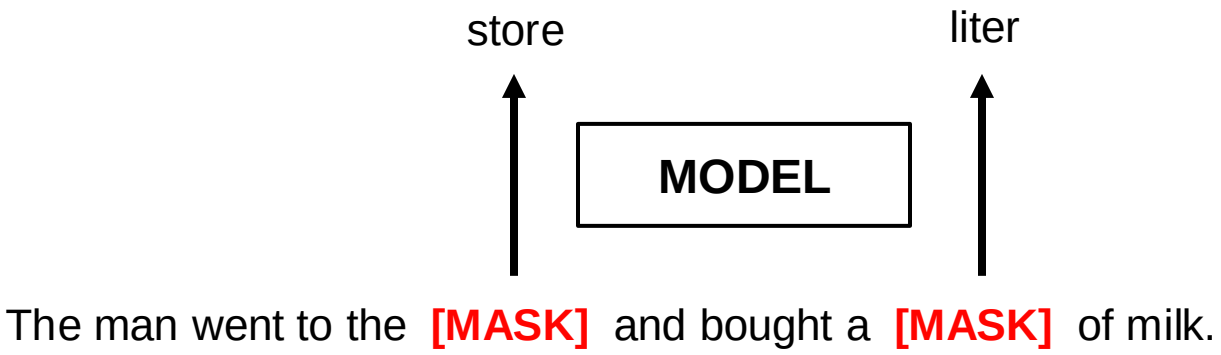


Jumper et al 2021

Masked Language Modeling in NLP

- **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**



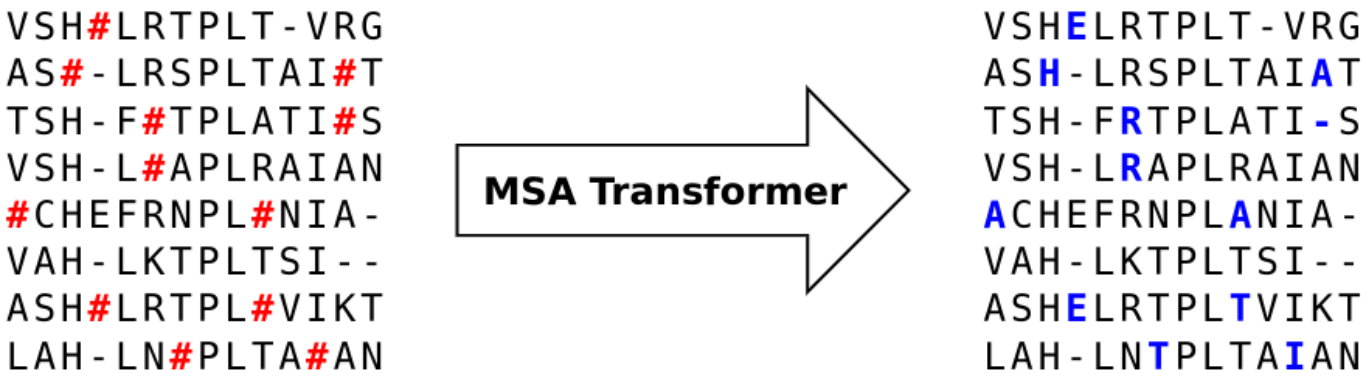
The model is trained to minimize a pseudo-likelihood loss:

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

MSA Transformer

- **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

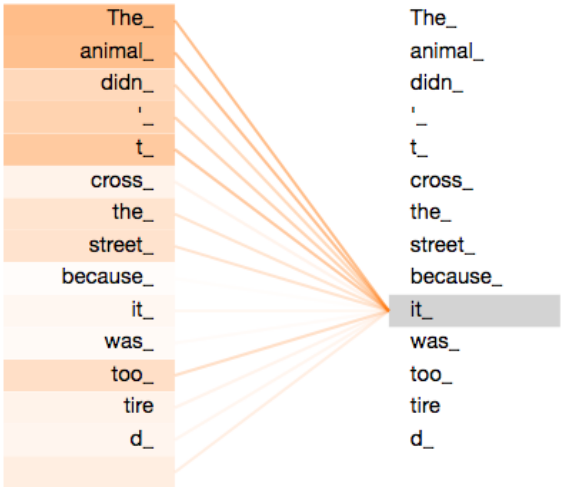
$\widetilde{\mathcal{M}}$ masked MSA

MSA Transformer is similar to AlphaFold's EvoFormer, but it is self-supervised
Here we focus on a model that works on MSAs – other ones work on single sequences

Architecture of MSA Transformer

- **Transformer architecture**

One attention head



L layers
A heads per layer

Full architecture

M tokens → *LA* matrices, each of size *M* × *M*

$BERT_{BASE}$: *L* = 12, *A* = 12
(Total parameters = 100M;
EvoFormer has 91M)

M tokens → *M* × *M* softmax values

The Illustrated Transformer, Alammar

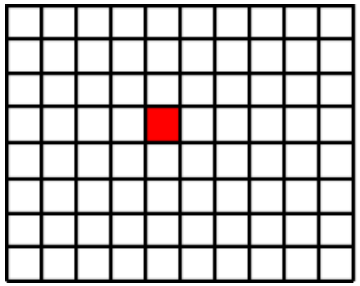
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021

$BERT_{BASE}$-like model with amino acids playing the part of words, trained with MLM objective

Relevant context for an amino acid?

Architecture of MSA Transformer

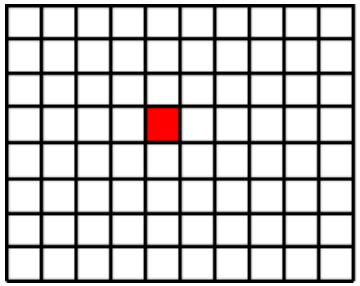
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



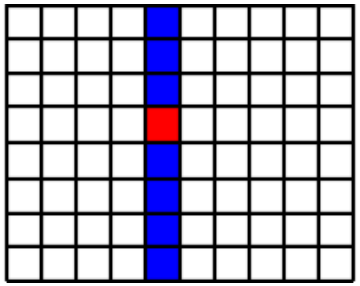
?

Architecture of MSA Transformer

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



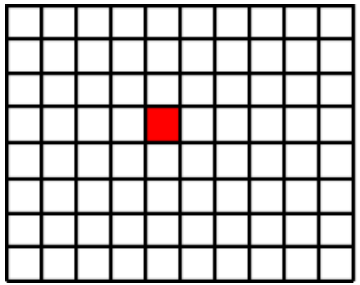
?



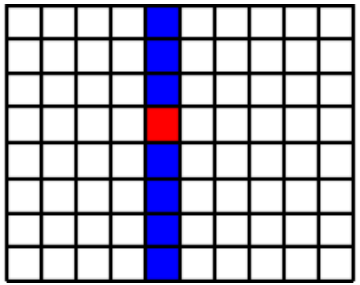
→ column attention

Architecture of MSA Transformer

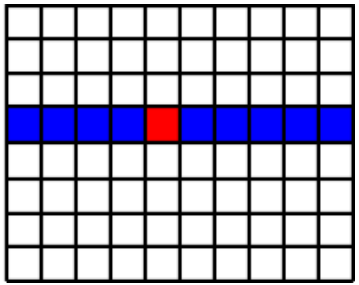
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



?



→ column attention

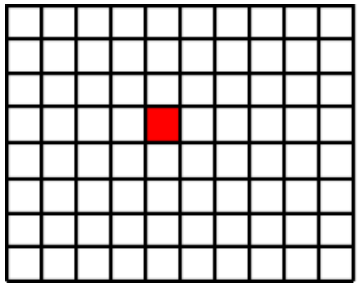


→ row attention

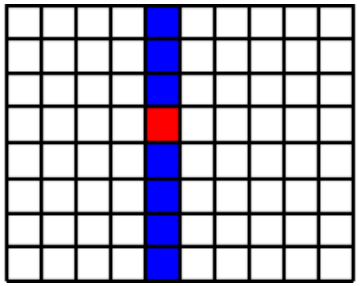
Context for an amino acid is both its column and its row ("axial attention" – Ho et al 2019)

Architecture of MSA Transformer

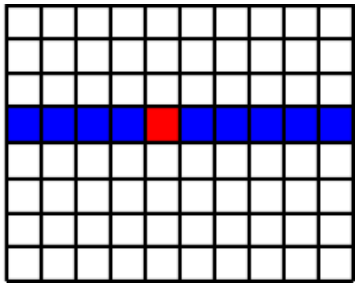
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



?



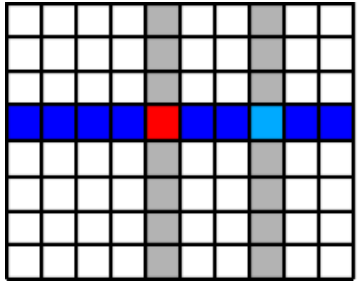
→ column attention

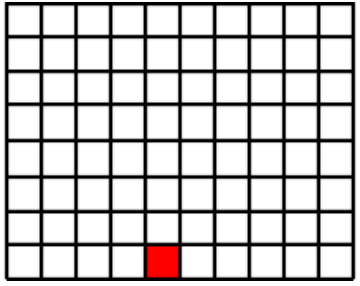


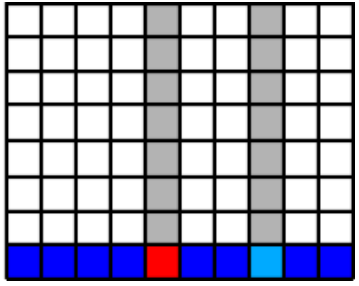
→ row attention

Context for an amino acid is both its column and its row ("axial attention" – Ho et al 2019)

Coevolution → row attention should be the same for all rows



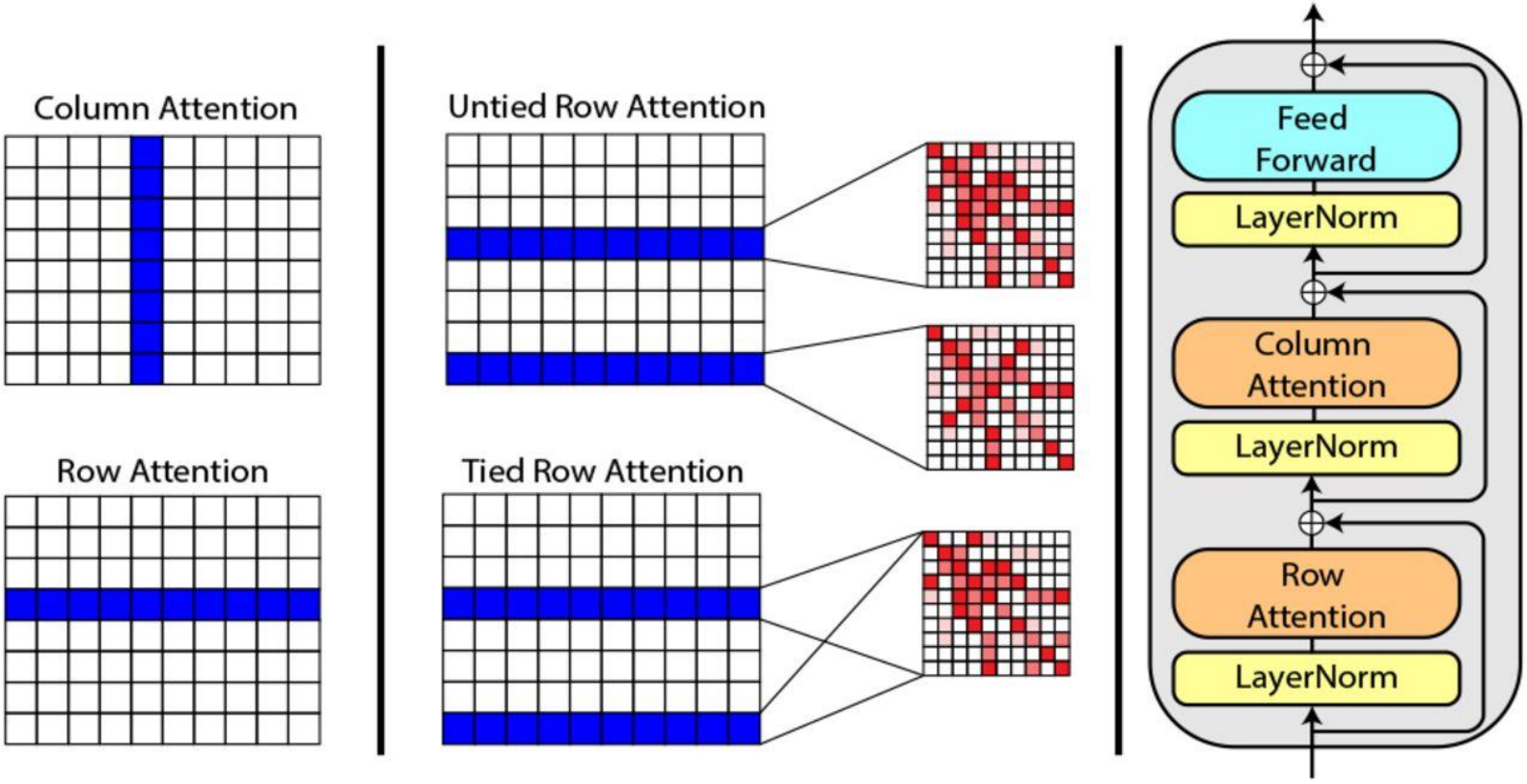




12 (layers) × 12 (heads) tied row attention units
12 × 12 independent column attention units
100M total parameters

Architecture of MSA Transformer

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

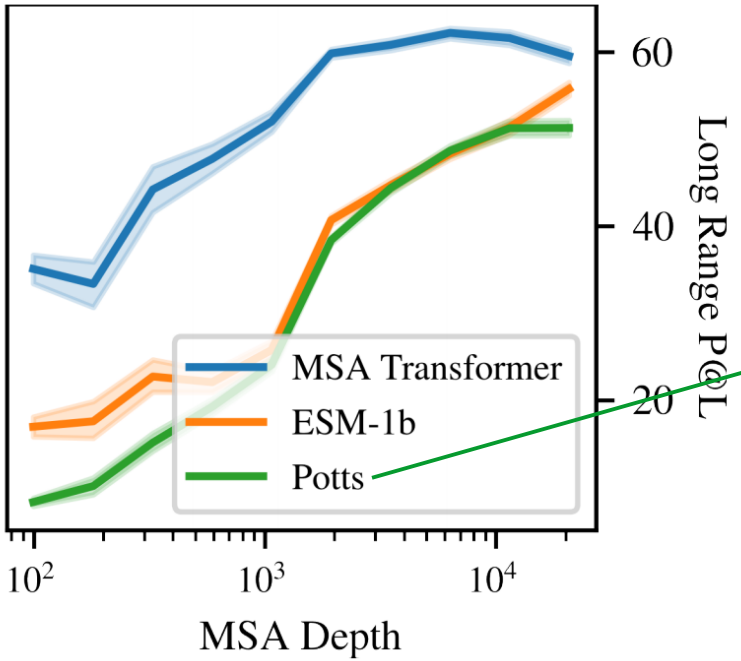


Training set:
- 26M MSAs corresponding to UniRef50 clusters
- average depth of MSAs: 1192

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]

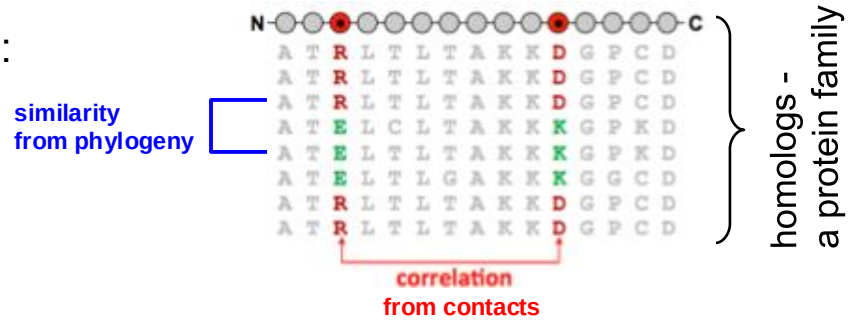
$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ - \left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\}$$

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

What kind of information is encoded in column attentions?

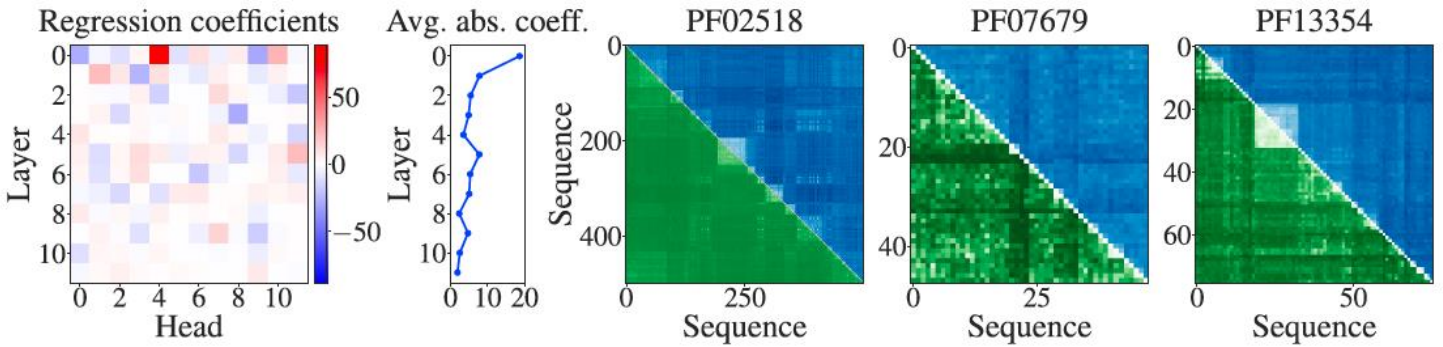
MSA Transformer's data representation

- **Column attentions encode phylogenetic relationships – Lupo, Sgarbossa & Bitbol 2022**
  - Motivation:



Casari et al 1995

- We fit a logistic model of the column attention matrices (averaged over columns) to predict the matrix of pairwise Hamming distances between sequences in MSAs
- Training: seed MSAs of 12 Pfam protein families; test: seed MSAs of 3 other Pfam families

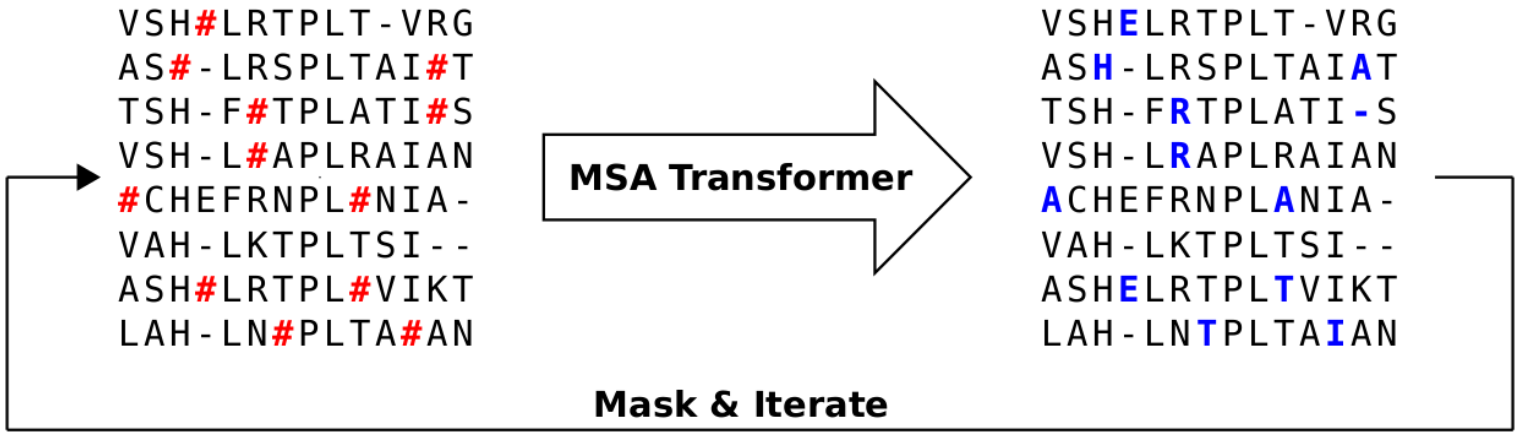


PF02518	PF07679	PF13354
R^2=0.60	R^2=0.28	R^2=0.67
ρ=0.90	ρ=0.85	ρ=0.92

→ A simple combination of column attention heads “implements” Hamming distance

Generating sequences with MSA Transformer

- **Iterative masking algorithm based on MLM – Sgarbossa, Lupo & Bitbol 2023 – PAPER**



Run iteratively this masking process on the same MSA → generate sequences

- Characterization of these sequences
- Comparison to sequences generated by a Potts model, using Metropolis-Hastings MCMC sampling (bmDCA Potts models are good generative models – Figliuzzi et al 2018, experimental validation Russ et al 2020)

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

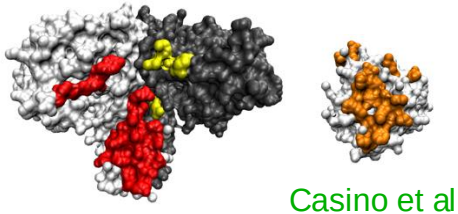
Pairwise maximum entropy model
Weigt, White et al 2009

one-body terms - fields

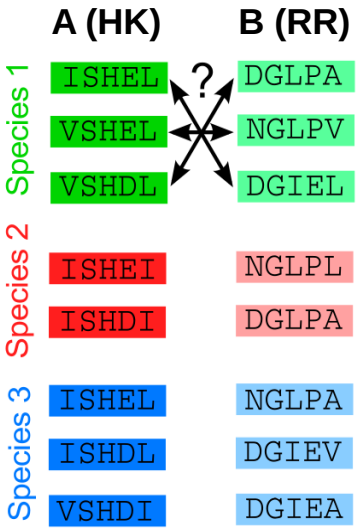
two-body terms - (direct) couplings

Predicting interaction partners

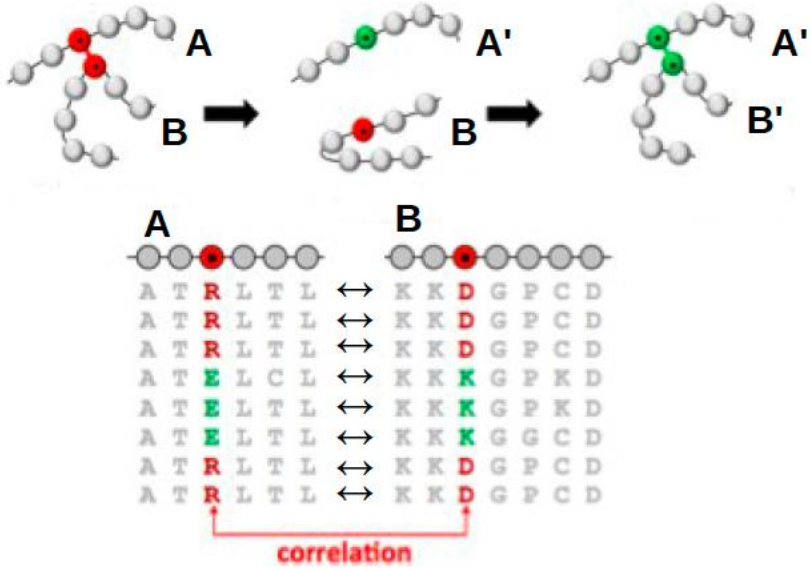
- **Coevolution can be used to infer interaction partners from sequences**



Casino et al. (2009)



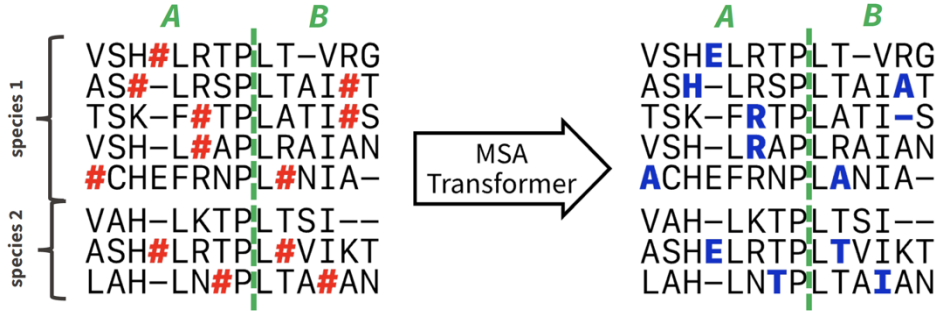
Within a species, which A interacts with which B?



→ Use correlations from coevolution to infer interaction partners (i.e. match paralogs):
Bayesian tree (Burger & van Nimwegen 2009),
Potts models (Bitbol et al 2016; Gueudre et al 2016)
Mutual Information (Bitbol 2018)
DCA or MI + phylogeny (Gandarilla-Pérez, Pinilla, Bitbol & Weigt 2023)

Inferring interaction partners using MSA Transformer

- **Paired MSAs as input for MSA Transformer – Lupo*, Sgarbossa* & Bitbol 2024 – PAPER**

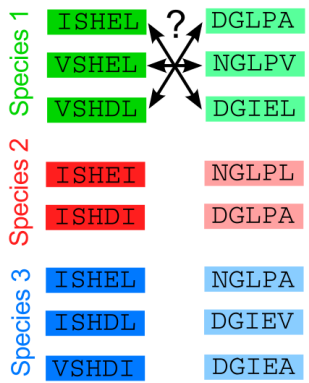


- MSA Transformer: trained on single-chain MSAs
- Take paired MSAs as input
- The MLM loss decreases as the fraction of correctly paired partners increases

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

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

- **Masked language modeling to infer interaction partners**



- **Goal:**
  In each species, find the permutation minimizing the MLM loss

- Permutation matrices approximated using the Sinkhorn operator (via a parameterization matrix)
  → Differentiable optimization problem
  Can be solved using gradient methods

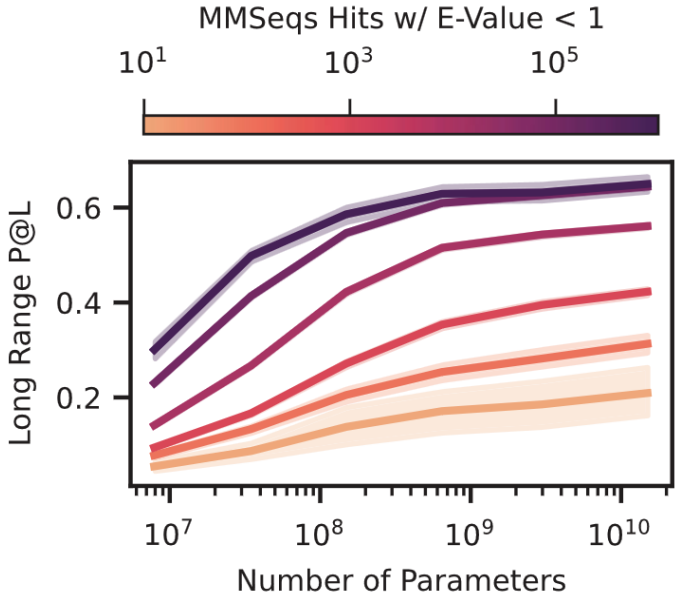
Are MSAs really necessary?

- **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 = 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 (Wua 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)
- strongly affected by the number ofexisting homologs!

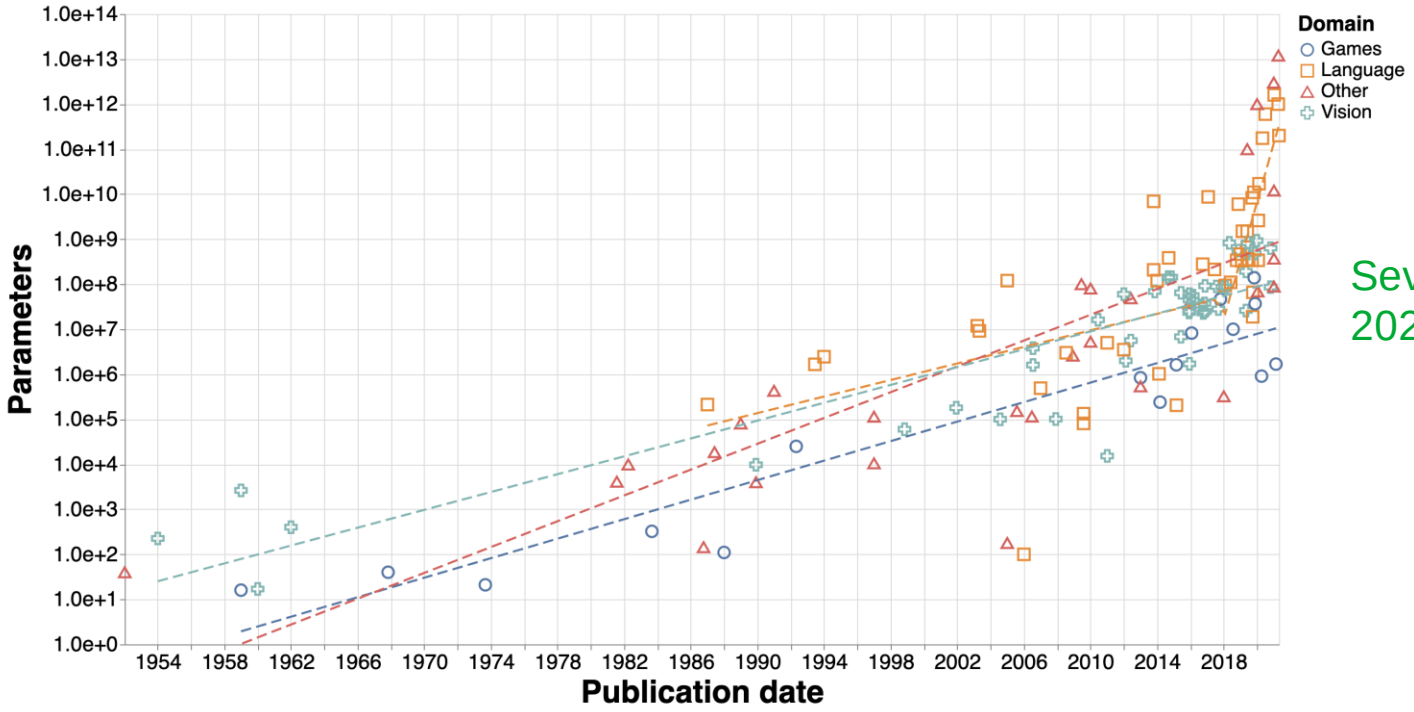
(Supervised) structure prediction:

- less good than AlphaFold2
- much faster → structure prediction at metagenomic scale

Are MSAs really necessary?

- **As of now, best performance for structure prediction requires MSAs**

Optimistic take for single-sequence LMs: we just need more parameters (Lin et al 2023)
"Our current models are very far from the limit of scale in parameters, sequence data, and computing power that can in principle be applied. We are optimistic that as we continue to scale, there will be further emergence. Our results showing the improvement in the modeling of low depth proteins point in this direction."

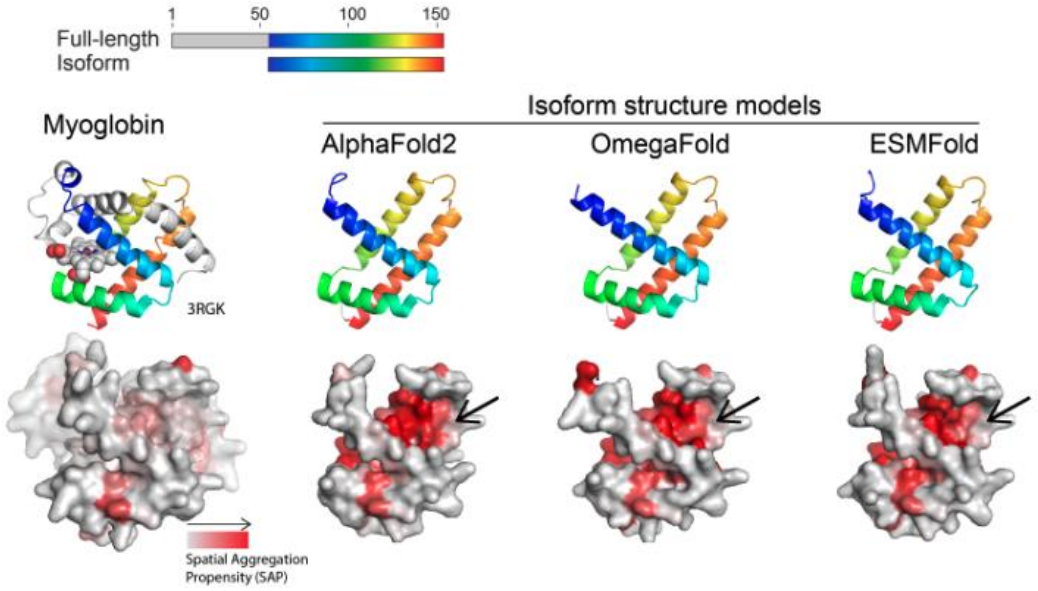


Sevilla et al, 2021

Are MSAs really necessary?

- **As of now, best performance for structure prediction requires MSAs**

Pessimistic take for single-sequence LMs: evolutionary information is crucial (Zhang et al 2024)
"Some have wondered if pLMs have finally solved the "protein folding problem", given their accurate structure prediction from single sequences and no supplied co-evolutionary signal in an input multiple sequence alignment. This should have been quickly debunked, as the accuracy of models was found to be highly correlated to the number of related proteins in the training set, indicating that the models store evolutionary information in their parameters"

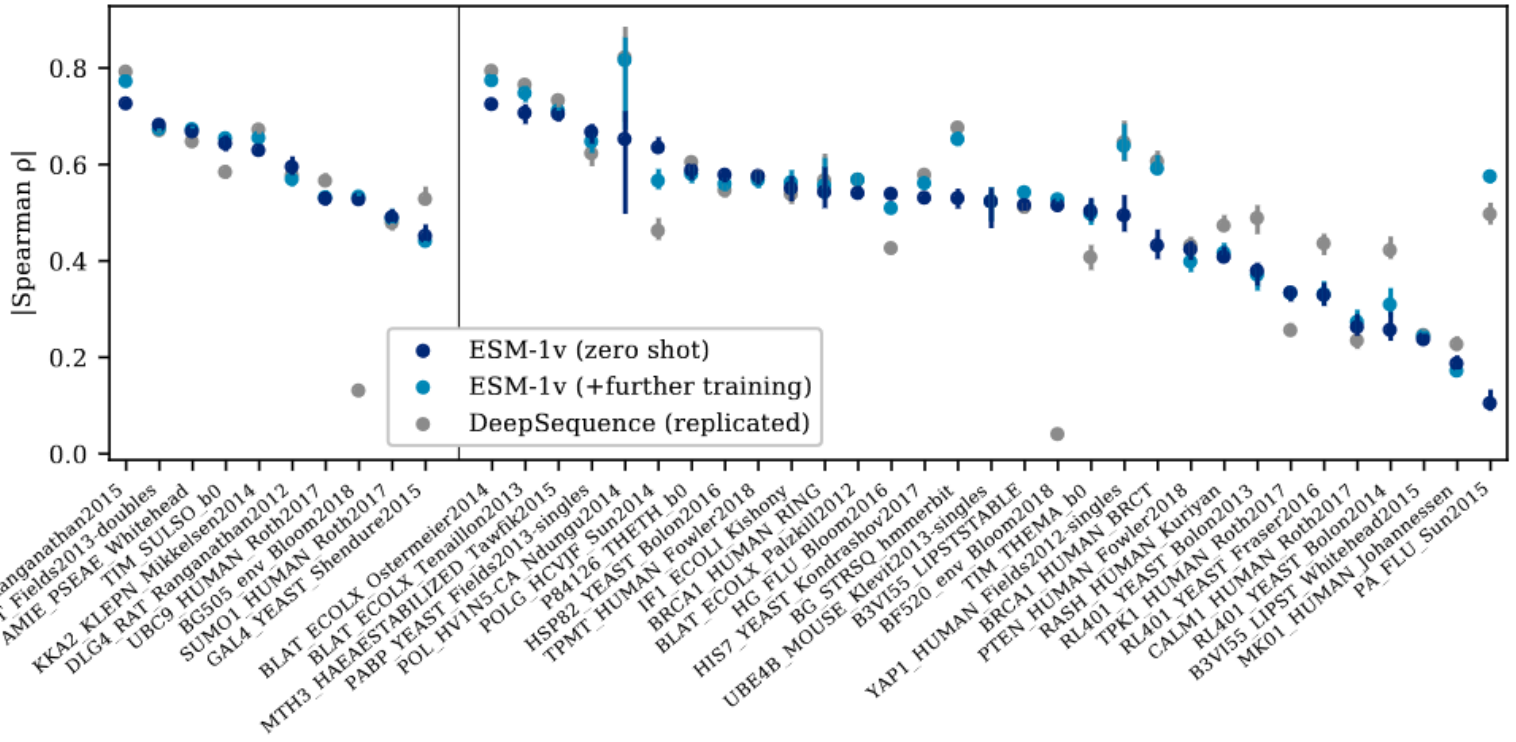


Isoform structure prediction is a challenge

Providing local windows of sequence information allows ESM-2 to best recover predicted contacts → pLMs
may predict contacts by storing motifs of pairwise contacts
(Zhang et al 2024)

Other applications of protein language models

- **Predicting the effect of mutations**

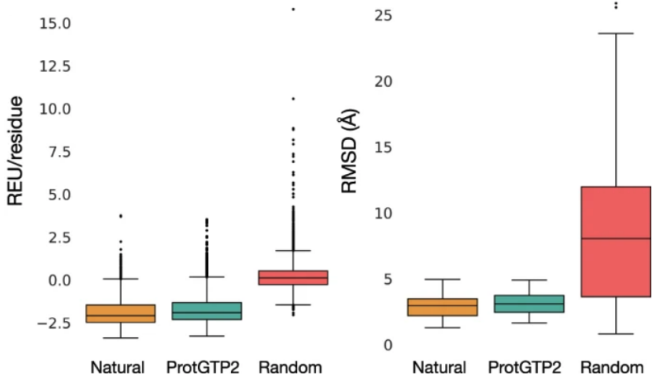


Ground truth: experimental deep mutational scans
Predictions: ESM-1v single-sequence protein language model (Meier et al 2021)

Other applications of protein language models

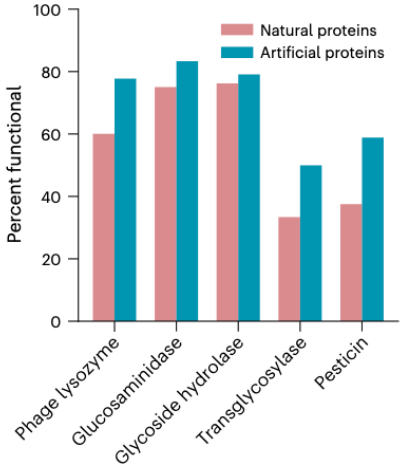
- **Designing new protein sequences**

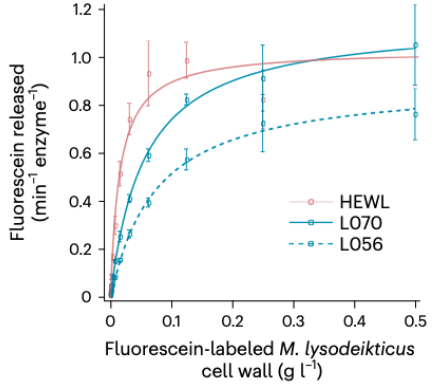
ProtGPT2 (Ferruz et al 2022): autoregressive transformer



Rosetta energy and flexibility patterns (from MD) similar to those of natural proteins

ProGen (Madani et al 2023): decoder transformer for *conditional* autoregressive generation – **PAPER**





Thanks!