Protein language models: generative properties and data representation

Anne-Florence Bitbol

EPFL SIB

Laboratory of Computational Biology and Theoretical Biophysics
Institute of Bioengineering, School of Life Sciences

AMLD, EPFL, *AI and the molecular world*
March 30, 2022

Outline

I. Context

1. Protein sequence data
2. Deep learning and protein sequence data

II. Generative properties of MSA Transformer

Damiano Sgarbossa, Umberto Lupo and Anne-Florence Bitbol

III. Phylogeny representation in MSA Transformer

Umberto Lupo, Damiano Sgarbossa and Anne-Florence Bitbol

Outline

I. Context

1. Protein sequence data
2. Deep learning and protein sequence data

II. Generative properties of MSA Transformer

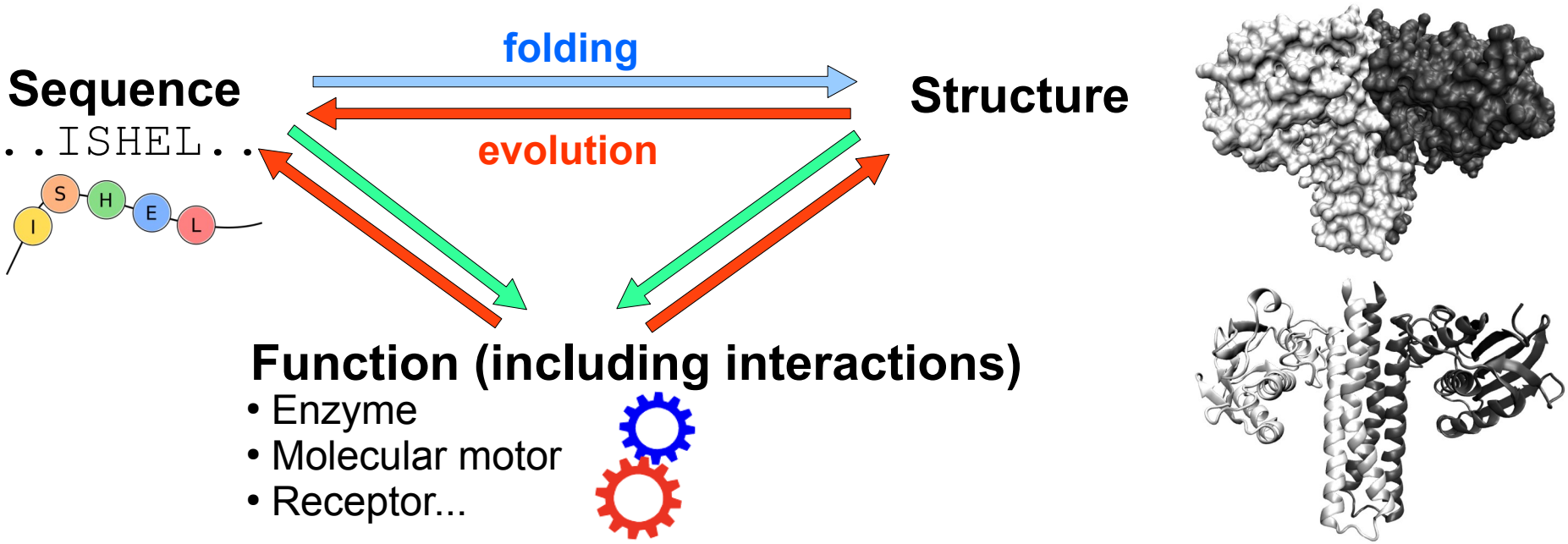
Damiano Sgarbossa, Umberto Lupo and Anne-Florence Bitbol

III. Phylogeny representation in MSA Transformer

Umberto Lupo, Damiano Sgarbossa and Anne-Florence Bitbol

Introduction: proteins

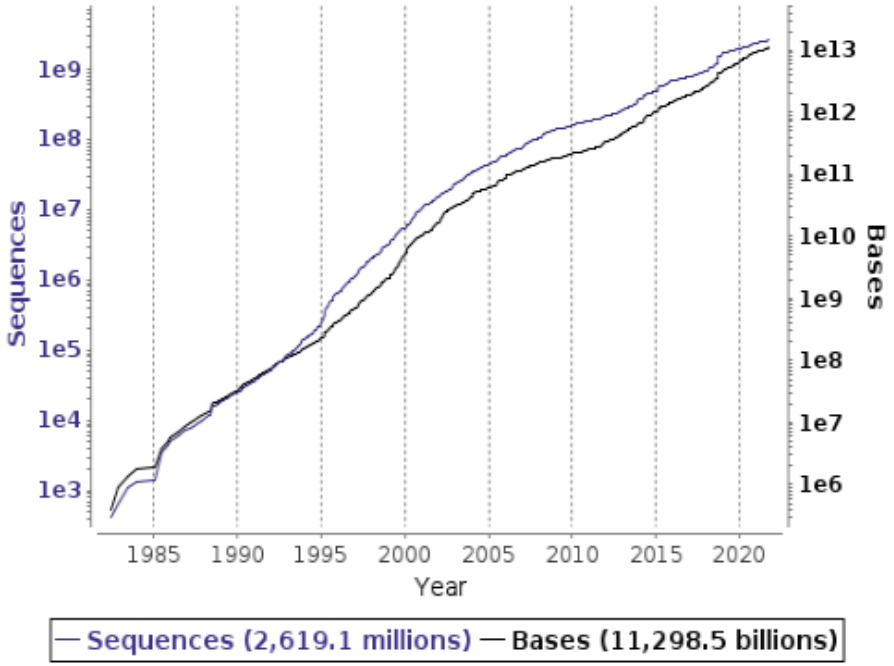
- **Understanding proteins**



- 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**

Protein sequence data

- **A growing amount of data**

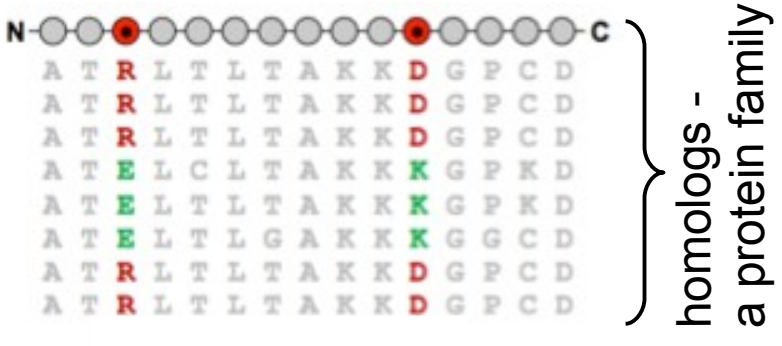


Accumulating sequence data
(currently > 10^9 sequences)

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

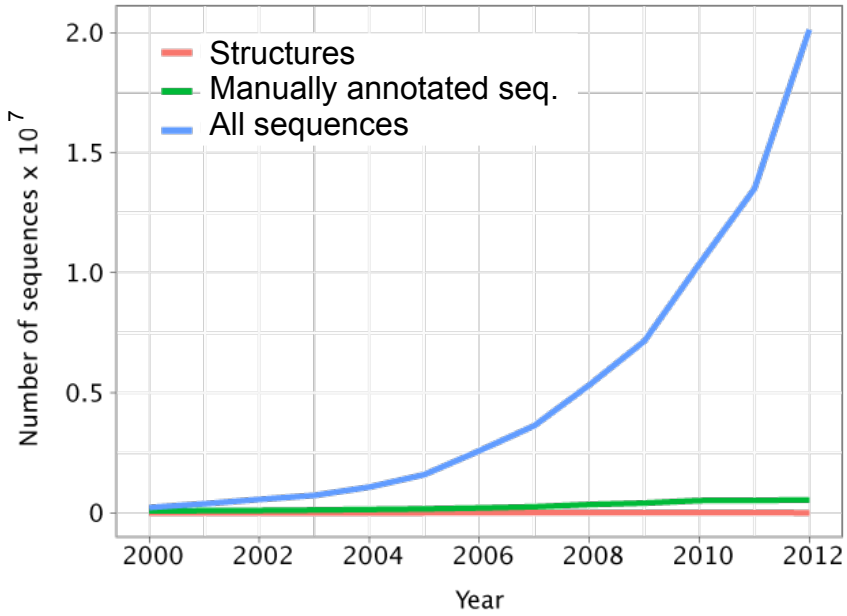
https://www.ebi.ac.uk/ena/browser/about/statistics

- **Protein families and multiple sequence alignments**



Protein sequence data

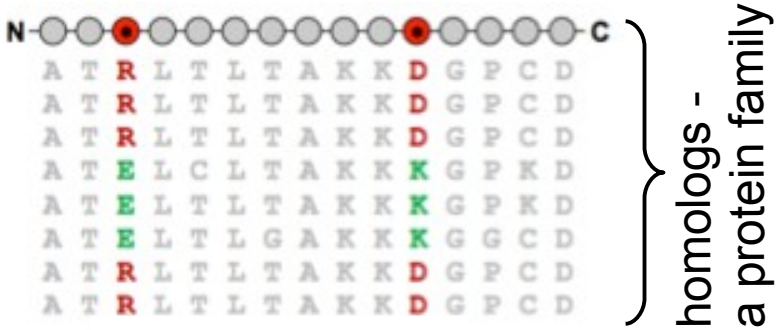
- **A growing amount of data**



Accumulating sequence data
(currently > 10^9 sequences)

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

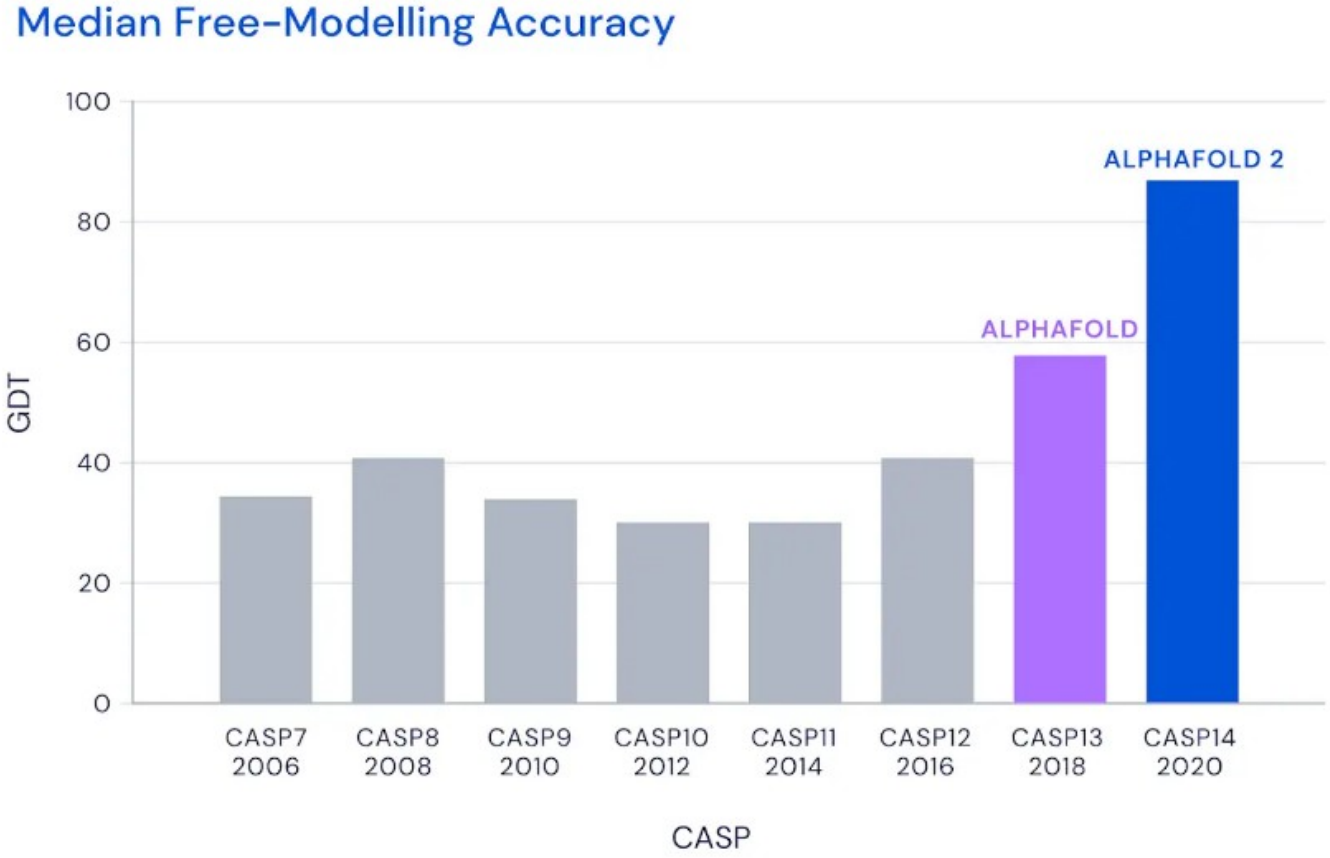
- **Protein families and multiple sequence alignments**



Deep learning and protein sequence data

- **Recent developments in protein structure prediction - Jumper et al 2021**

  - Deep learning approaches – AlphaFold, AlphaFold2 – won CASP13 and **CASP14**
  - **Supervised** models to predict structure (uses experimental structures as input)
  - AlphaFold2 uses natural language processing methods:
    Attention (Bahdanau et al 2014), transformer architecture (Vaswani et al 2017)



GDT: global distance test

median score: 92.4 GDT
→ RMSD ~ 1.6 A

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

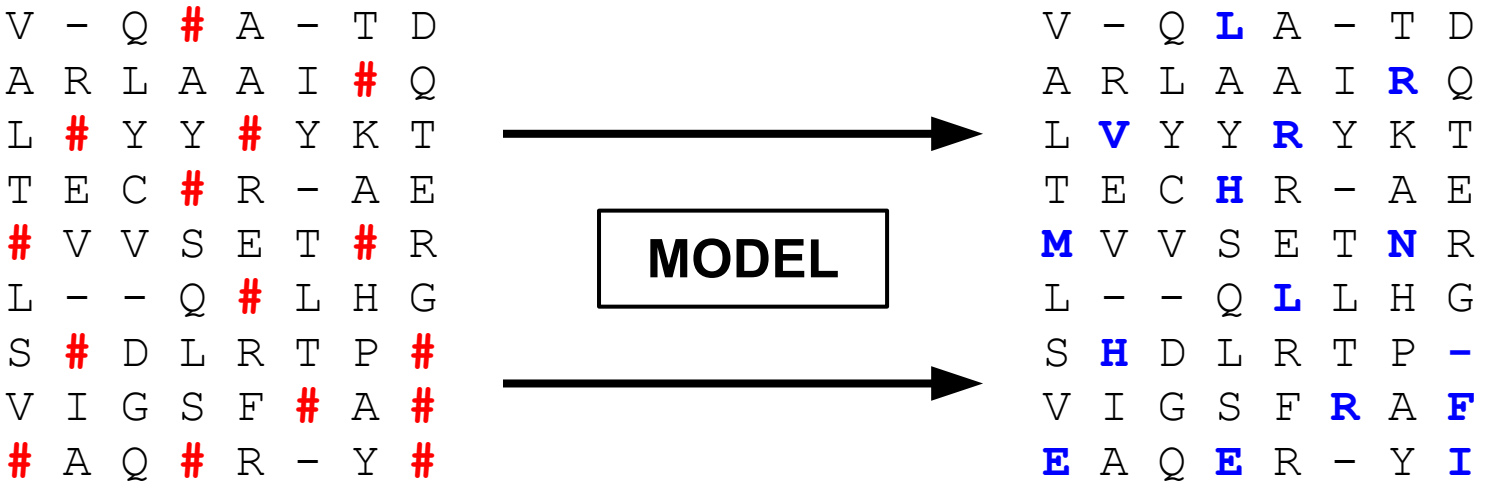
- **How about non-supervised learning using NLP-based architectures?**

Masked Language Modeling for protein sequence data

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

Randomly **mask** (**#**) a fraction of the **residues** and train the model to predict them, using the surrounding **context** (generative task)

Note: Here we focus on a model that works on MSAs – multiple other ones work on single sequences, ignoring protein families and MSAs



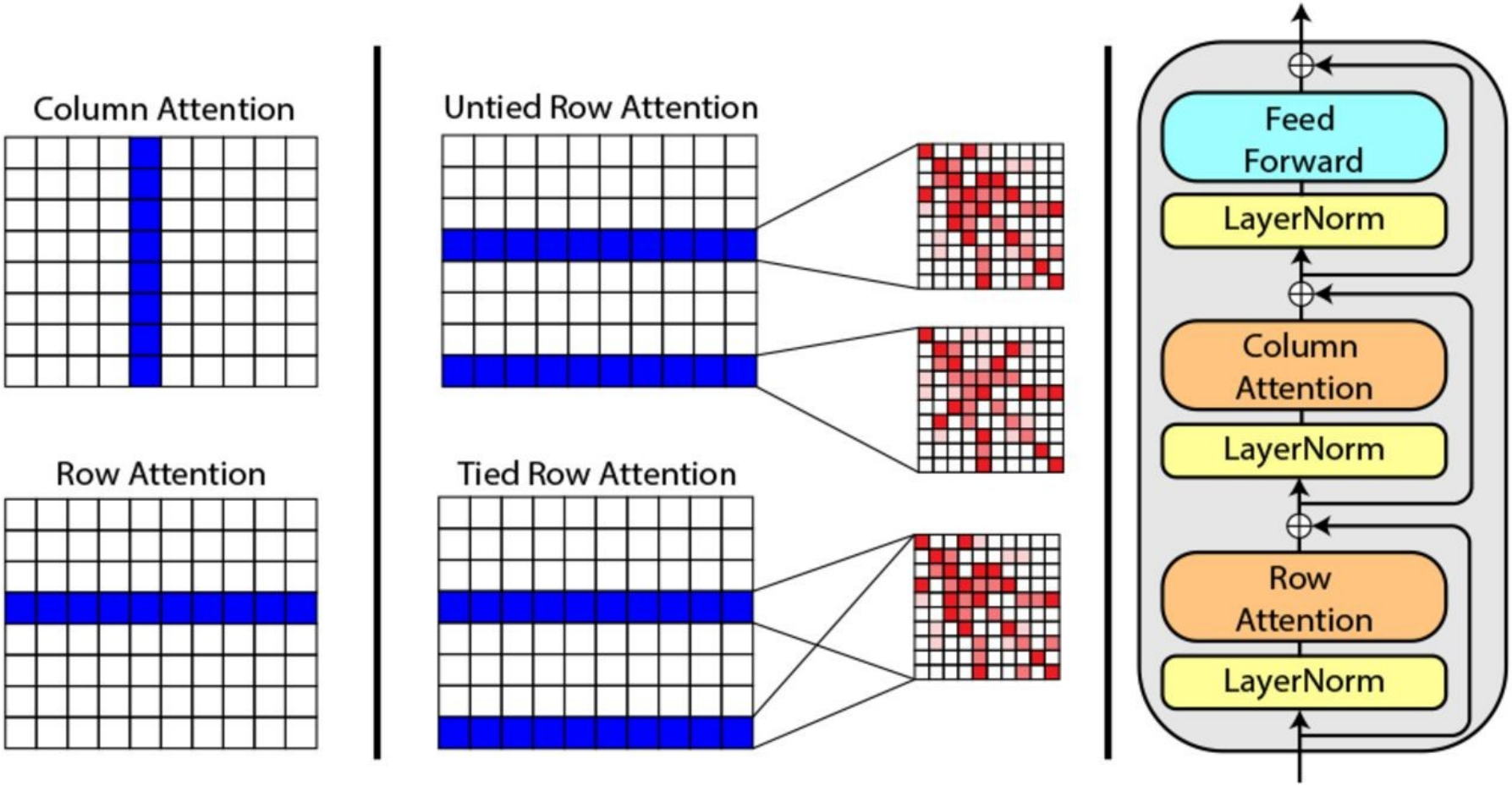
The model is trained to minimize a pseudo-likelihood loss:

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

MSA Transformer – Rao et al 2021

- **Transformer architecture adapted to the 2D structure of protein MSAs**

$BERT_{BASE}$-like model; **one-letter amino acids** as tokens; trained with MLM objective
Context for an amino acid = its **row** and **column** ("**axial attention**" – Ho et al 2019)

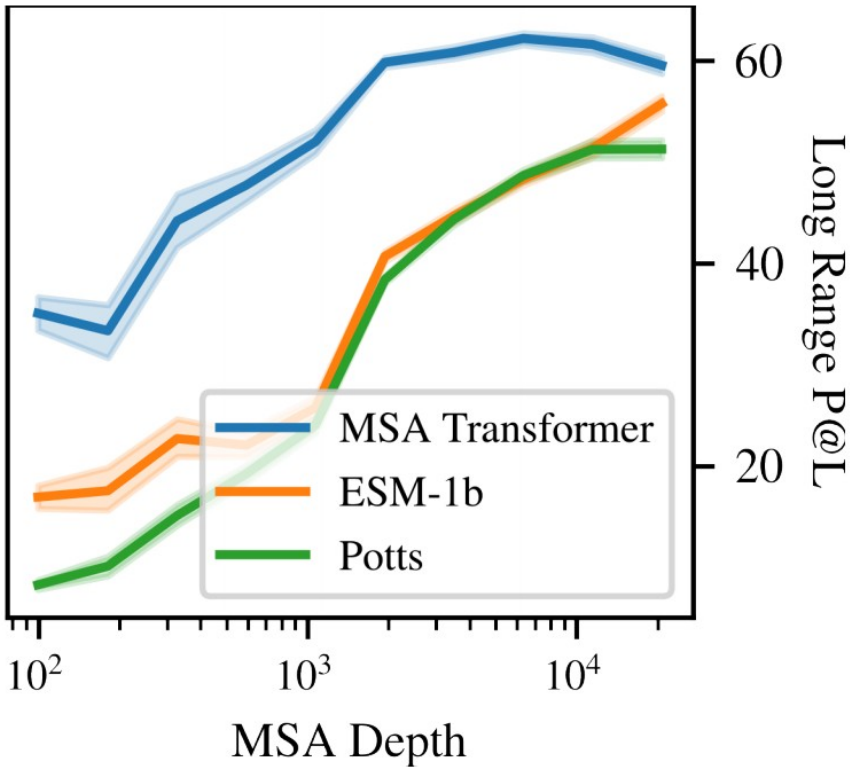


Rao et al 2021

Unsupervised structural contact prediction by MSA Transformer

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

  • Coevolution → to predict a masked amino acid, we should "attend to" **highly correlated positions** in the same row, and these are often in contact in the 3D structure
  • Indeed, simple combinations of the **row attention** softmax matrices (summed over all input rows) allow **contact prediction**:



Contact prediction performance

MSA Transformer outperforms both BERT-like single-sequence models (ESM-1b) and Potts model at the unsupervised contact prediction task

Outline

I. Context

1. Protein sequence data
2. Deep learning and protein sequence data

II. Generative properties of MSA Transformer

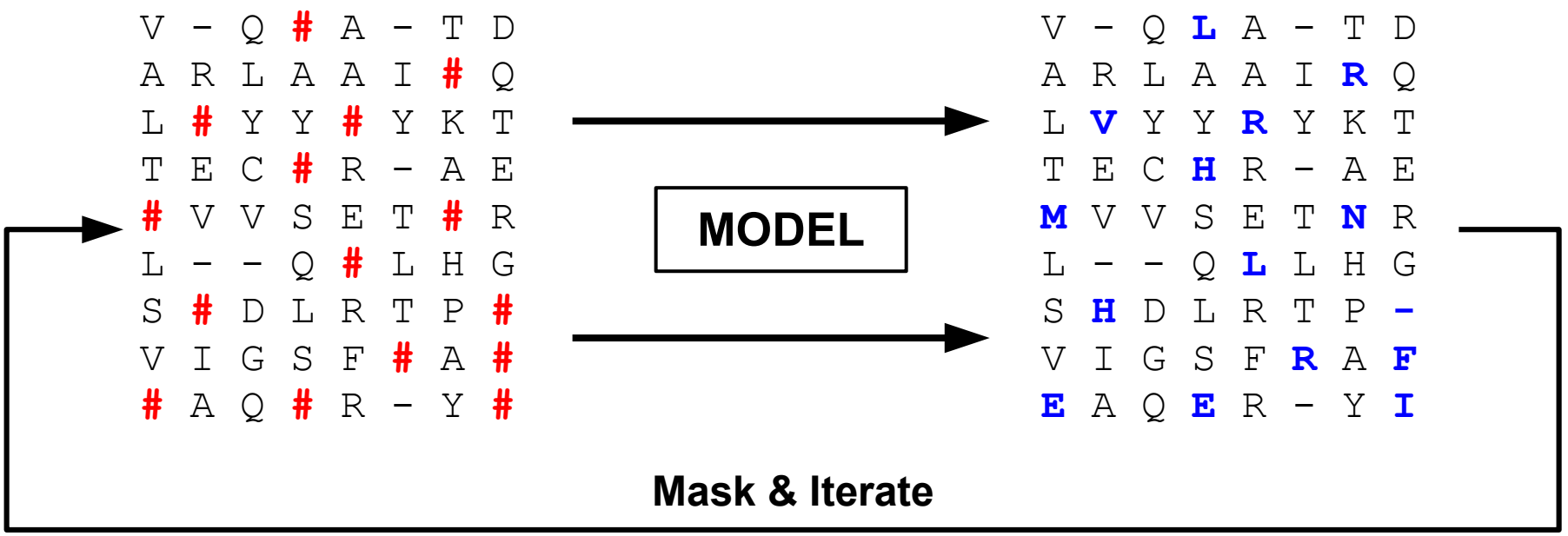
Damiano Sgarbossa, Umberto Lupo and Anne-Florence Bitbol

III. Phylogeny representation in MSA Transformer

Umberto Lupo, Damiano Sgarbossa and Anne-Florence Bitbol

Generating sequences with MSA Transformer

- **Iterative masking algorithm based on MLM**



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, Russ et al 2020)

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

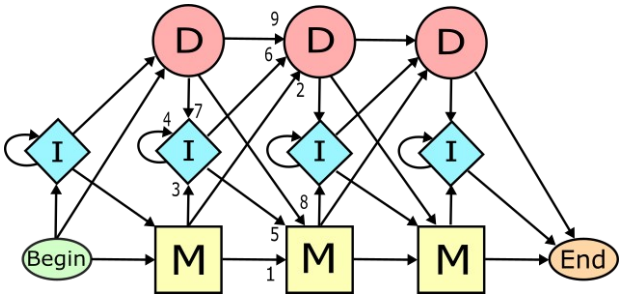
Pairwise maximum entropy model
Weigt, White et al 2009

one-body terms - fields

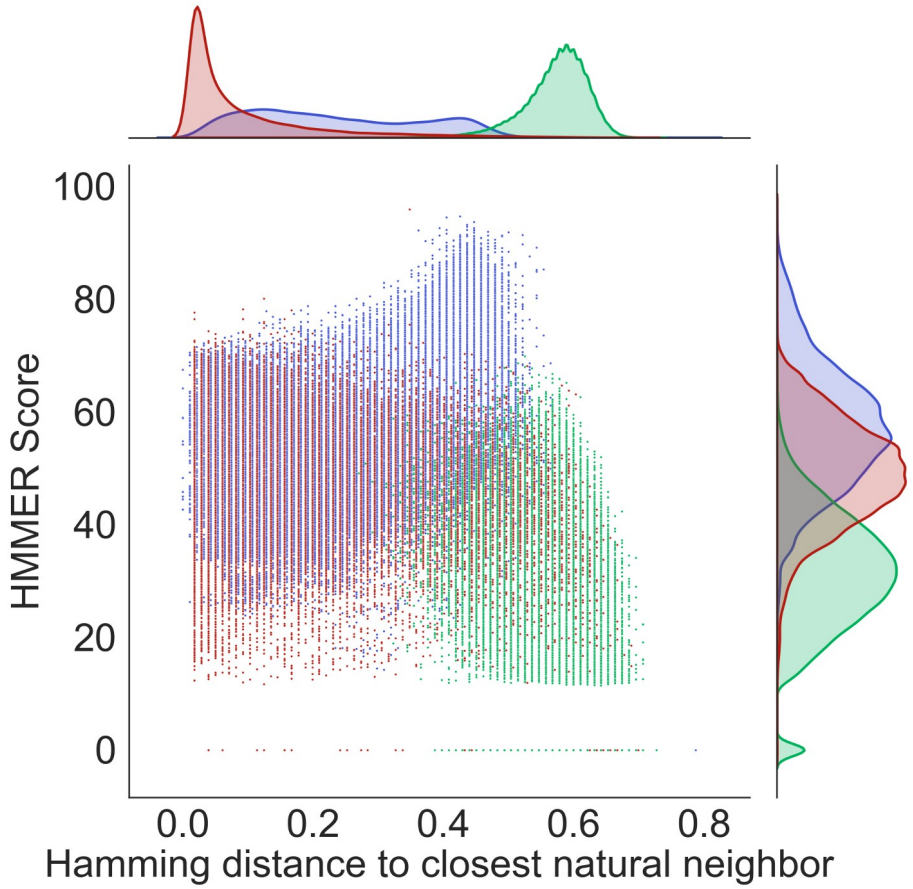
two-body terms - (direct) couplings

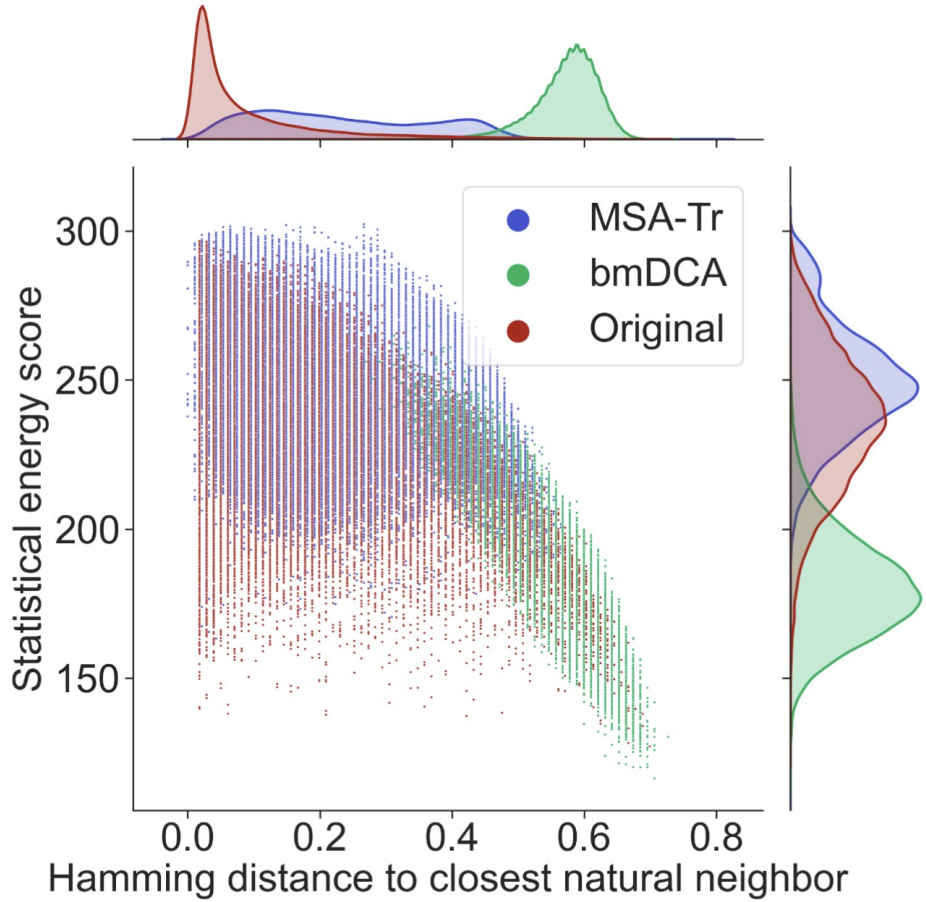
Characterization of sequences generated by MSA Transformer

- **Scores of generated sequences vs. distance to closest natural sequence**



$$-H = \sum_{i,j} e_{i,j}(\alpha_i, \alpha_j) + \sum_i h_i(\alpha_i)$$

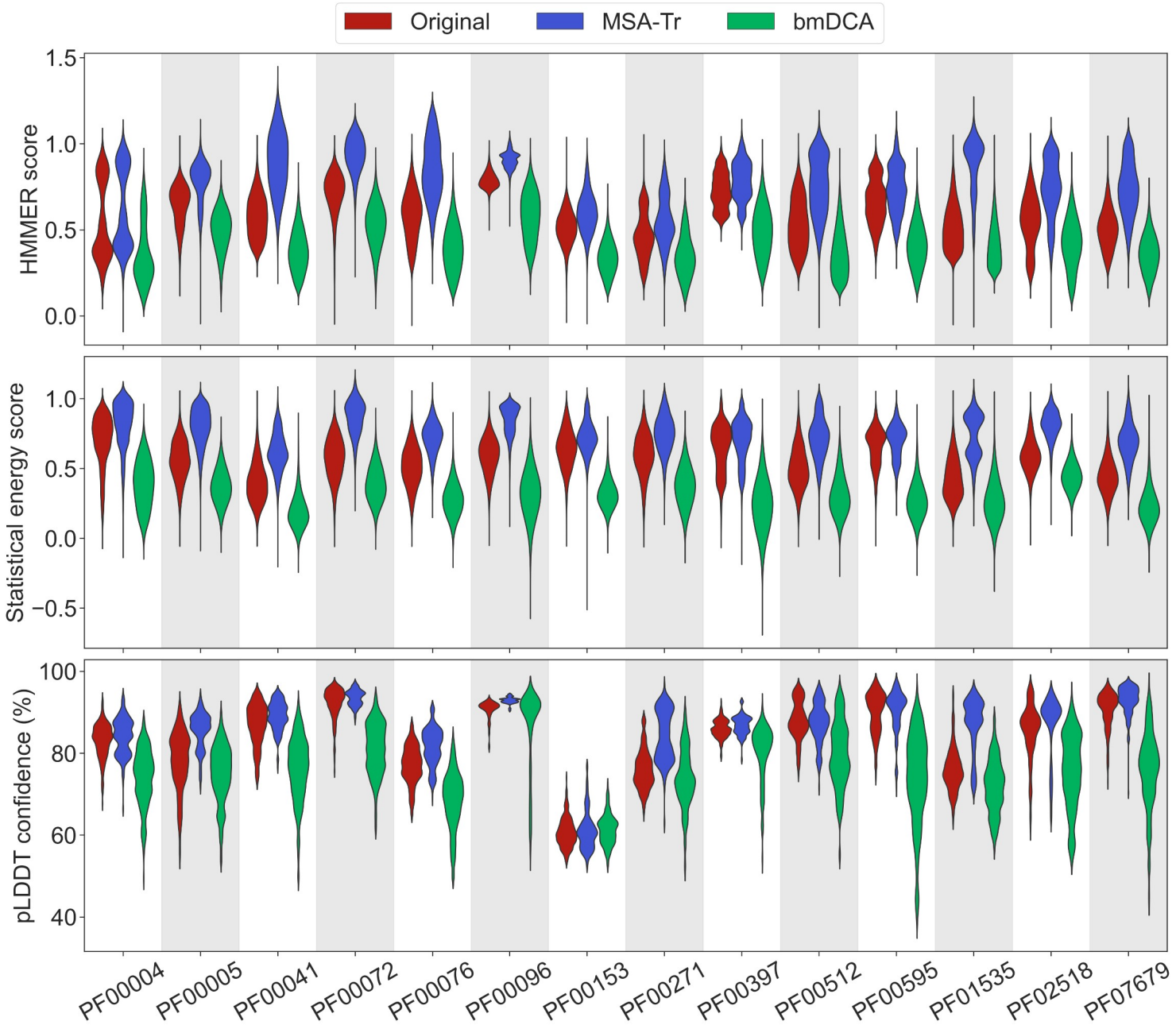




Protein family: PF00153 – mitochondrial ADP/ATP carrier

Characterization of sequences generated by MSA Transformer

- **Distributions of three scores for 14 different deep Pfam protein families**



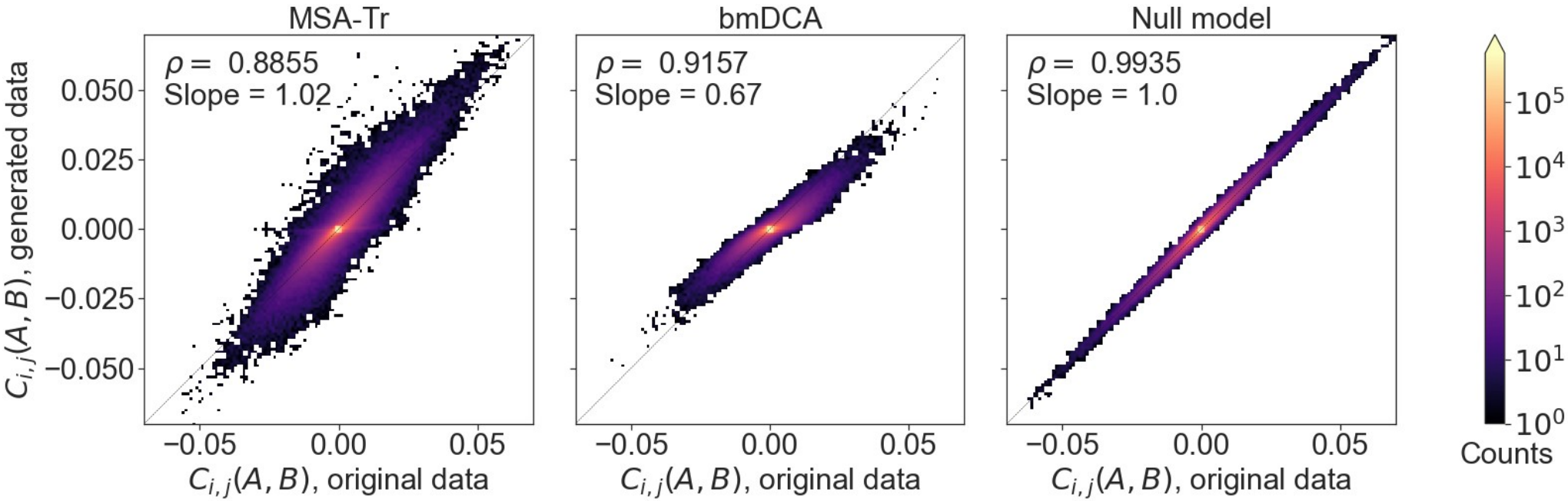
Characterization of sequences generated by MSA Transformer

- **Second order statistics for PF00153 (mitochondrial ADP/ATP carrier)**

We consider different statistics and information measures, starting from:

$f_i(A)$; $f_{ij}(A,B)$; $f_{ijk}(A,B,C)$ = 1, 2 and 3 point frequencies

First, consider second order connected correlations: $c_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B)$

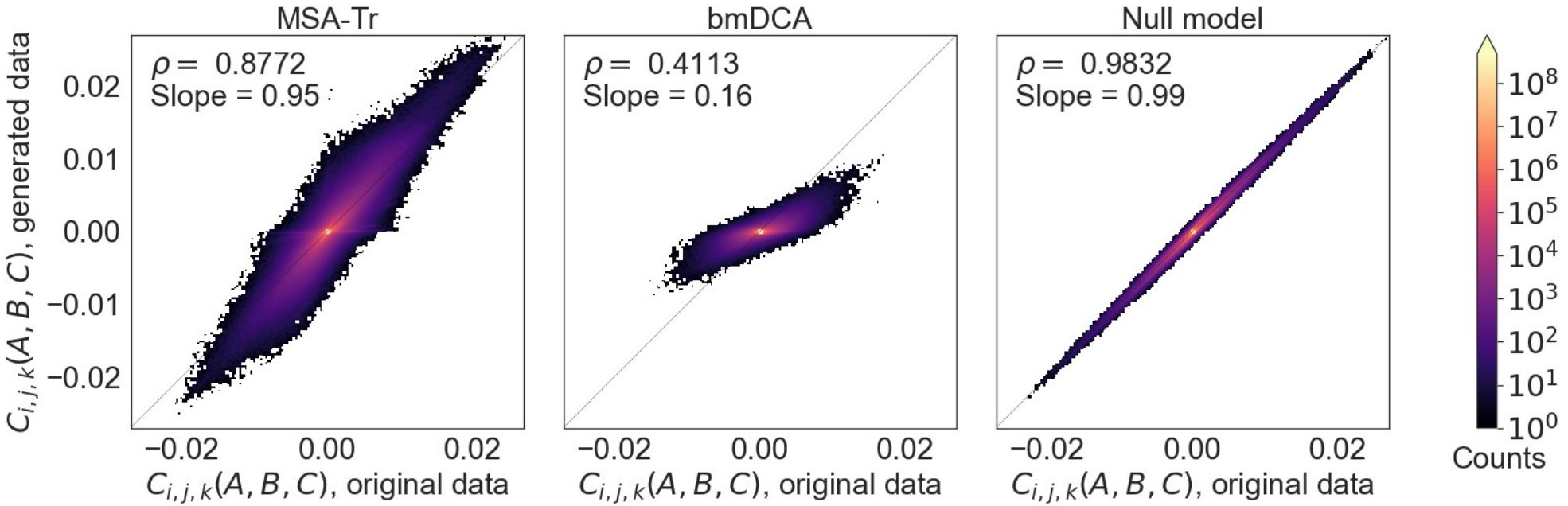


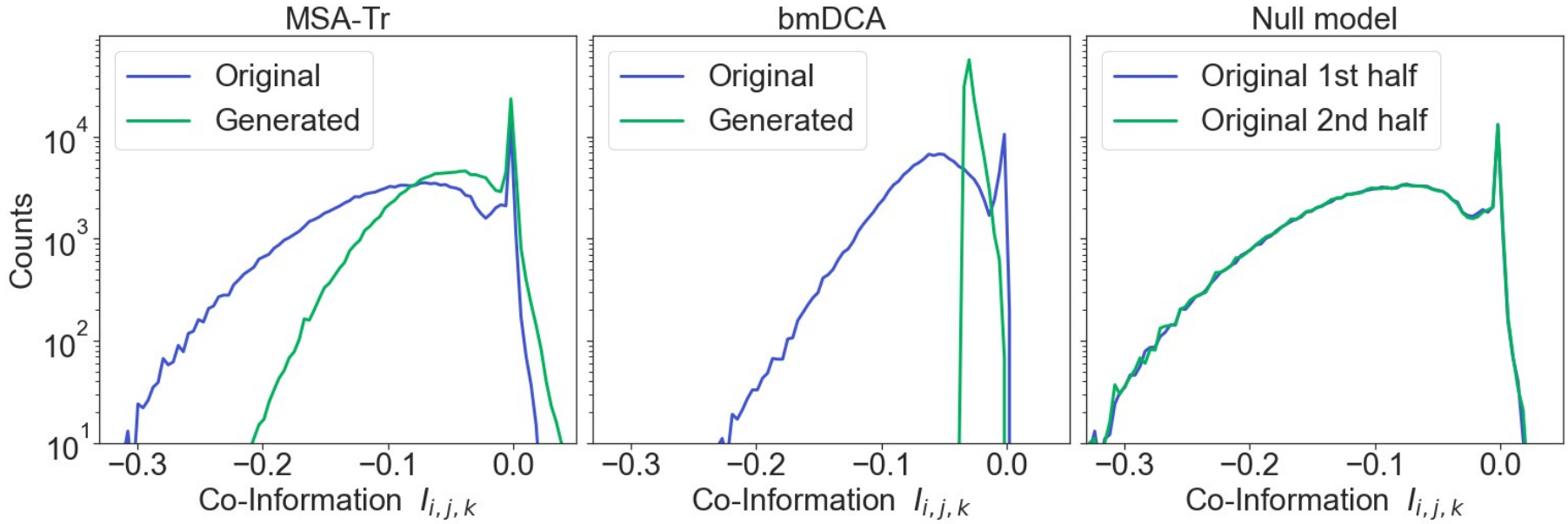
For this measure, bmDCA (Potts model) often performs better than MSA Transformer, esp. for the Pearson correlation (on 14 deep Pfam protein families)

Recall that Potts models are trained to reproduce one- and two-body statistics

Characterization of sequences generated by MSA Transformer

- **Third order statistics and co-information for PF00153 (mitochondrial ADP/ATP carrier)**

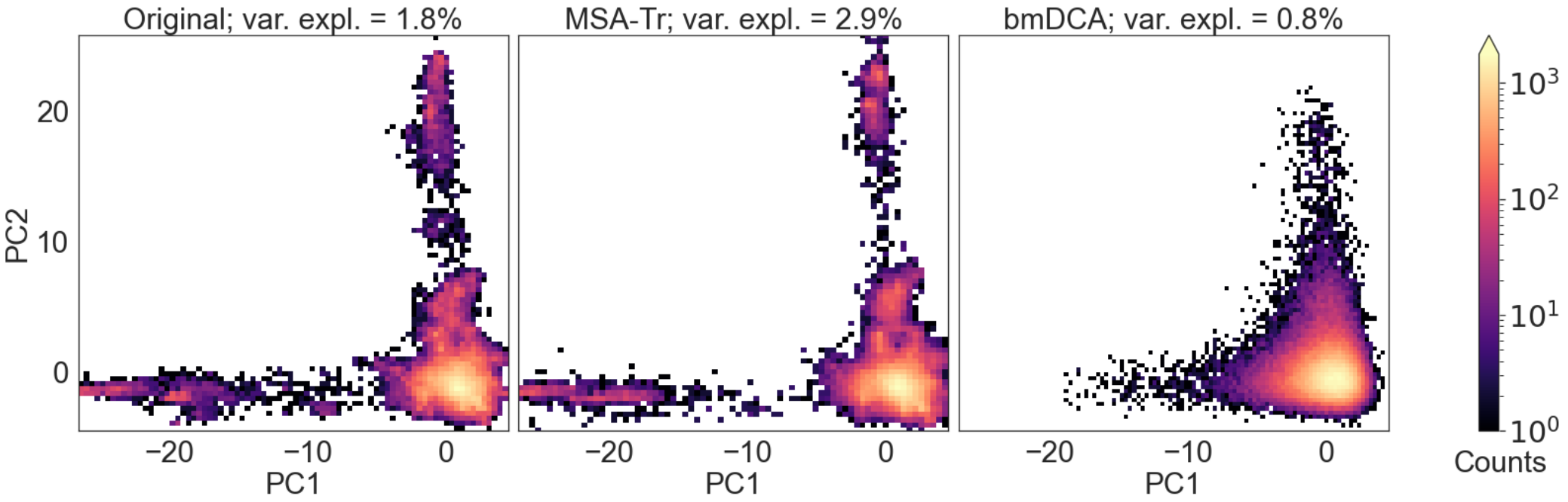




Characterization of sequences generated by MSA Transformer

- **Distribution of sequences in sequence space – PF00153 (mitochondrial ADP/ATP carrier)**

We perform PCA on sequences and consider the top two PCs



Sequence

A F F - Q - L L A G A

One-hot encoding

100000...0 000010...0 ...

Conclusion

- **Summary**

  - Method to generate sequences from MSA Transformer exploiting the MLM objective
  - Generated sequences have good scores – overall even better than natural ones
  - One- and two-body statistics are better reproduced by bmDCA than by MSA Transformer, but the opposite holds for three-body statistics
  - MSA Transformer is a good candidate to generate novel proteins from a protein family

- **Perspective**

  - Protein design

- **Preprint**

  Coming soon!

Outline

I. Context

1. Protein sequence data
2. Deep learning and protein sequence data

II. Generative properties of MSA Transformer

Damiano Sgarbossa, Umberto Lupo and Anne-Florence Bitbol

III. Phylogeny representation in MSA Transformer

Umberto Lupo, Damiano Sgarbossa and Anne-Florence Bitbol

Introduction

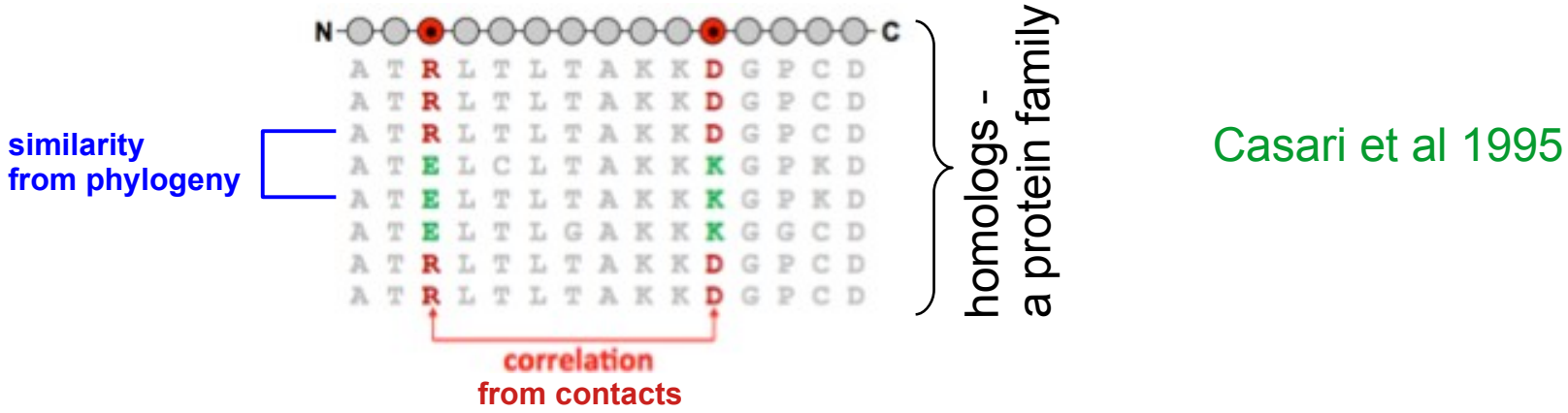
- **Motivation and question**

MSA Transformer uses axial attention (Ho et al 2019) and computes both row and column attention matrices

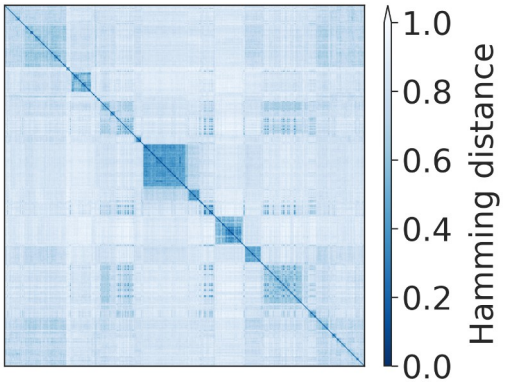
Tied row attentions: capture structural contacts – Rao et al 2021

Column attentions: their average across columns and then across the "first *M*" correlate (somewhat) with phylogenetic weights (1/(number of neighbors))
→ more attention to more diverse sequences – Rao et al 2021

→ Do column attention matrices encode phylogenetic relationships?

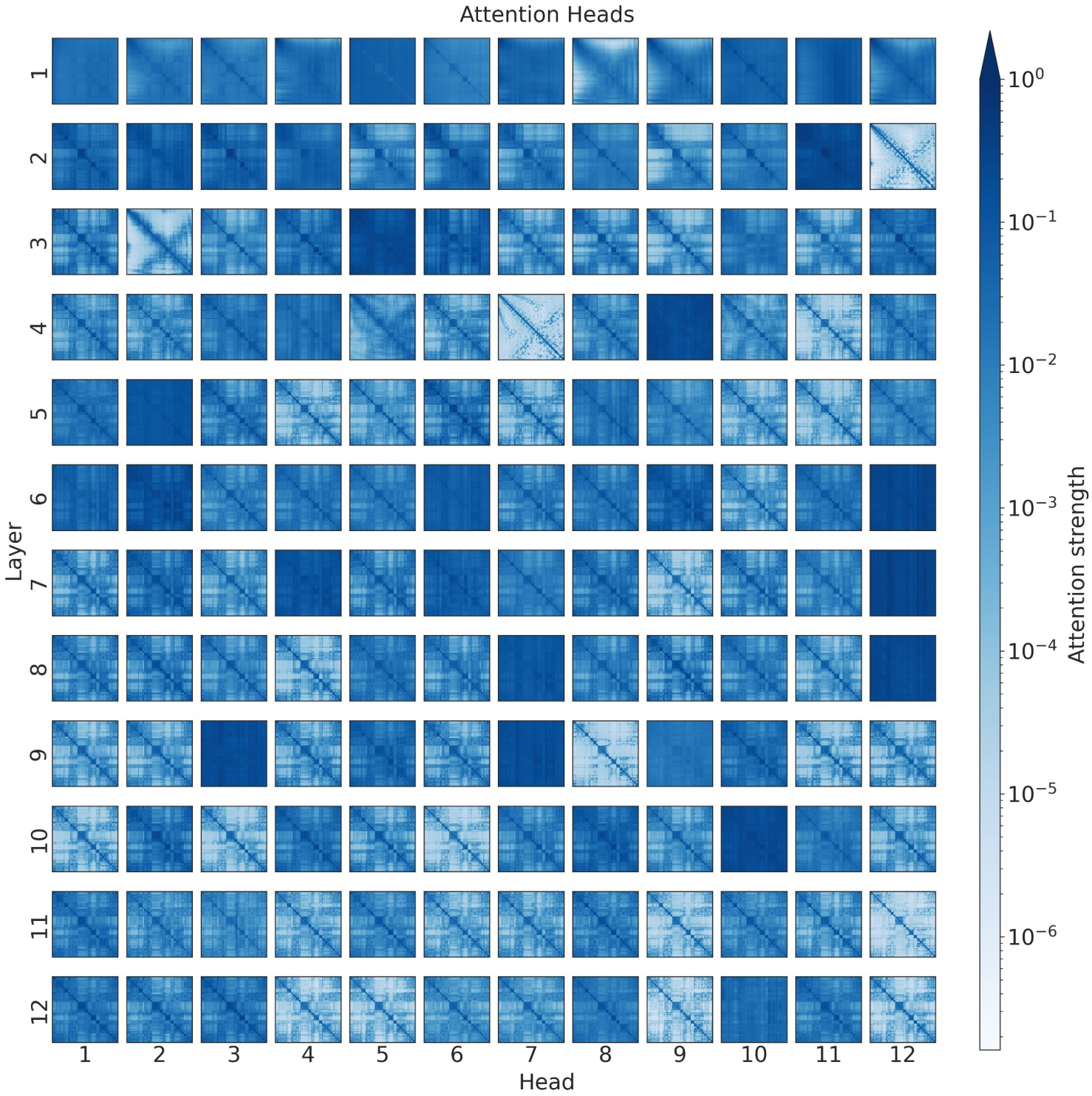


Casari et al 1995



Visual comparison of the Hamming distance matrix and of column attention matrices in MSA Transformer, averaged over columns

(for protein family PF02518, HATPase_c domain, seed MSA)



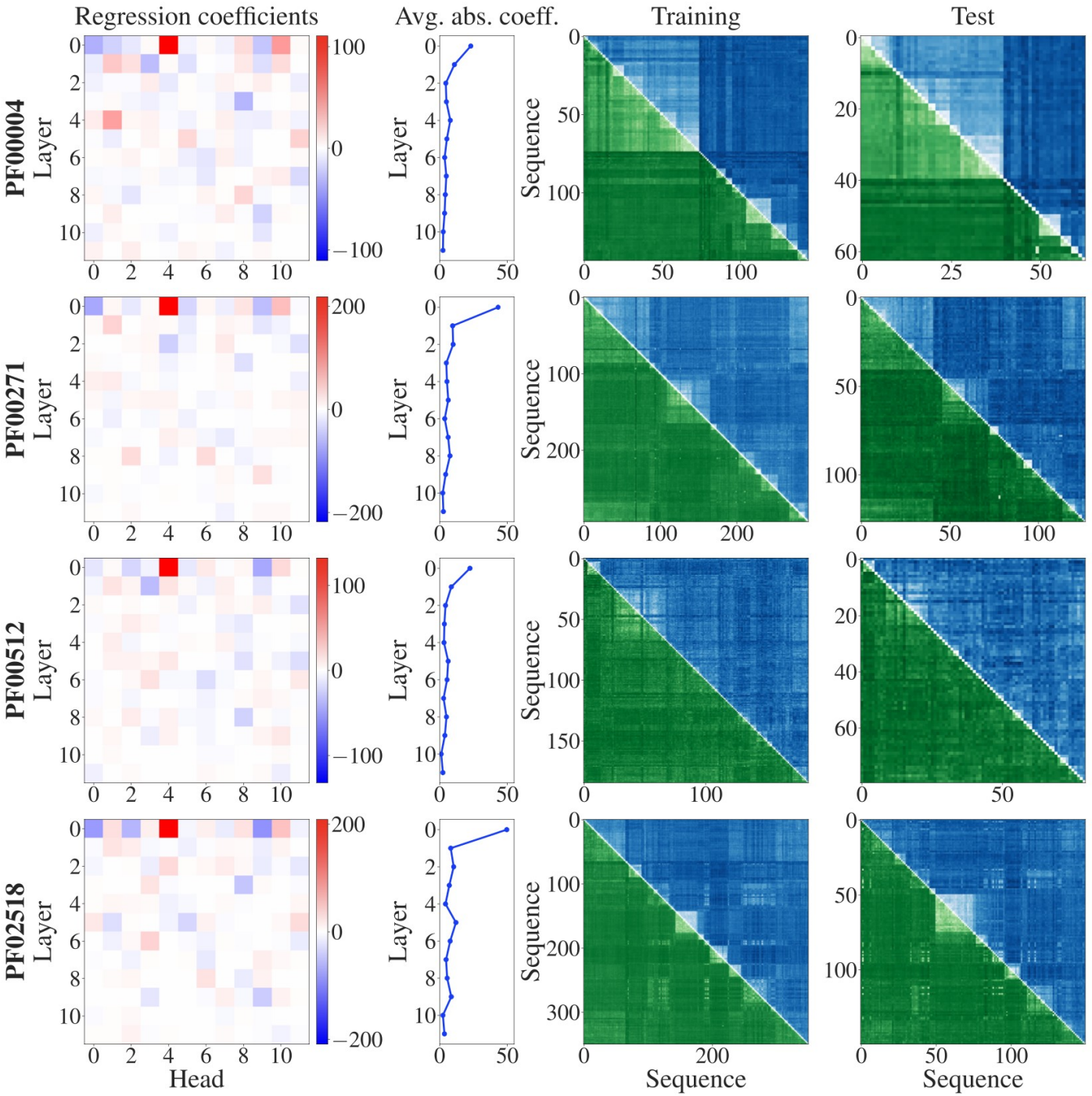
Predicting Hamming distances from column attentions

Regression to predict Hamming distances (HD) within a family

For an MSA (Pfam seed):

- Average column attentions across columns
- Fit a **logistic model** to regress **HD matrix** entries from entries in the attention matrices
- **Test** on held-out portion of the HD matrix

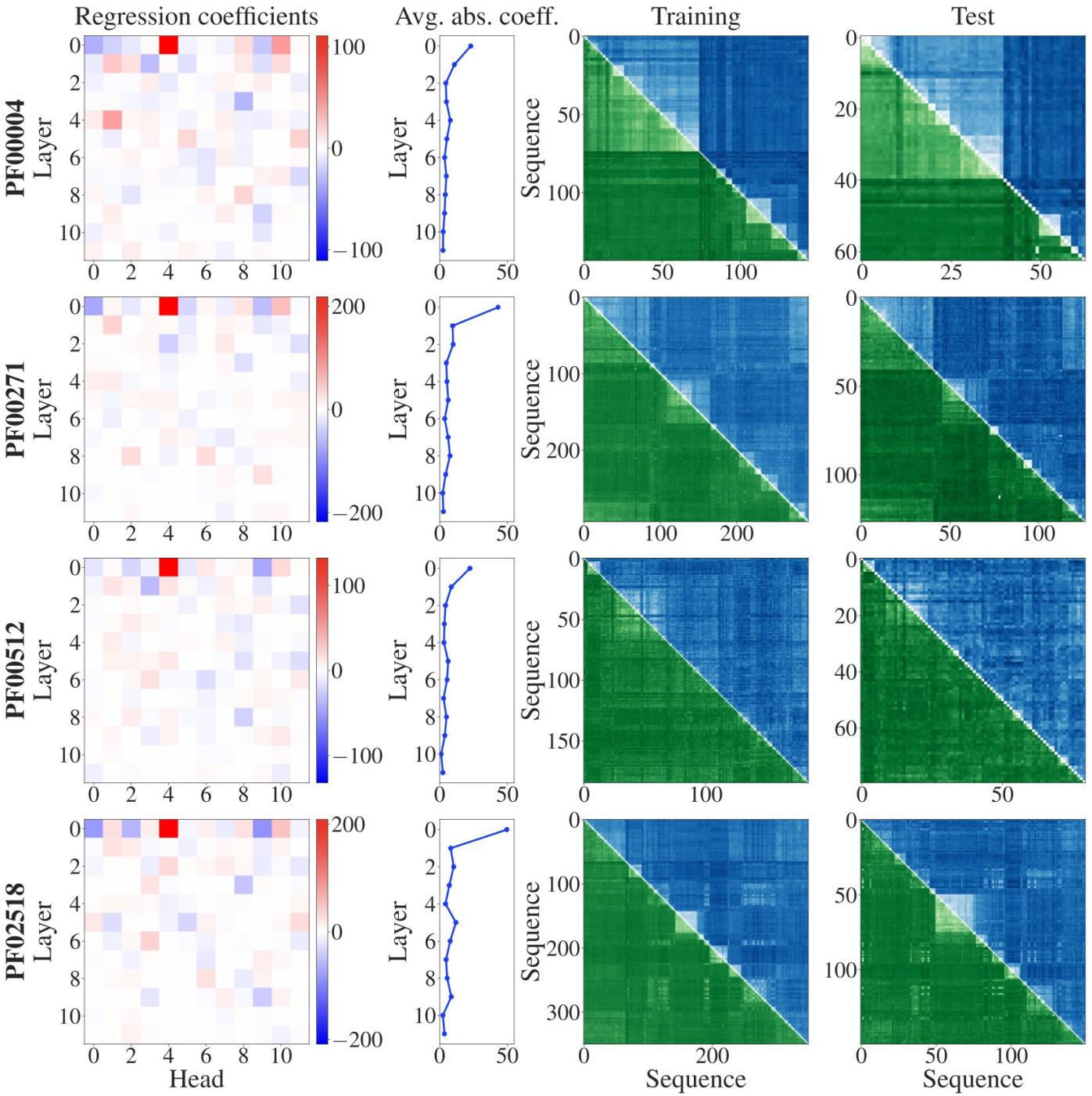
- Larger regression coefficients in **early layers**
- Some heads fire **consistently** across different MSAs



Predicting Hamming distances from column attentions

Regression to predict Hamming distances (HD) within a family

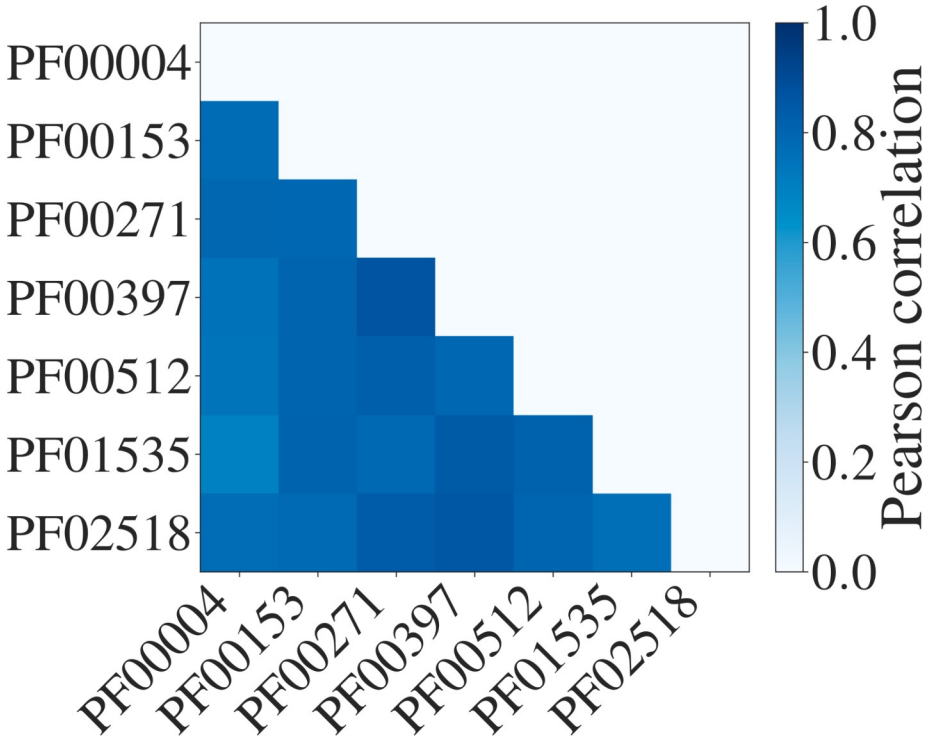
Family	R^2
PF00004	0.97
PF00005	0.99
PF00041	0.98
PF00072	0.99
PF00076	0.98
PF00096	0.94
PF00153	0.95
PF00271	0.94
PF00397	0.84
PF00512	0.94
PF00595	0.98
PF01535	0.86
PF02518	0.92
PF07679	0.99
PF13354	0.99



Predicting Hamming distances from column attentions

- **Universality?**

Is there a **universal** simple combination of column attention heads which “implements” Hamming distance?



The regression coefficients found in the MSA-specific regressions are highly correlated amongst MSAs

(mean/min/max Pearson correlations: 0.80/0.69/0.87)

Predicting Hamming distances from column attentions

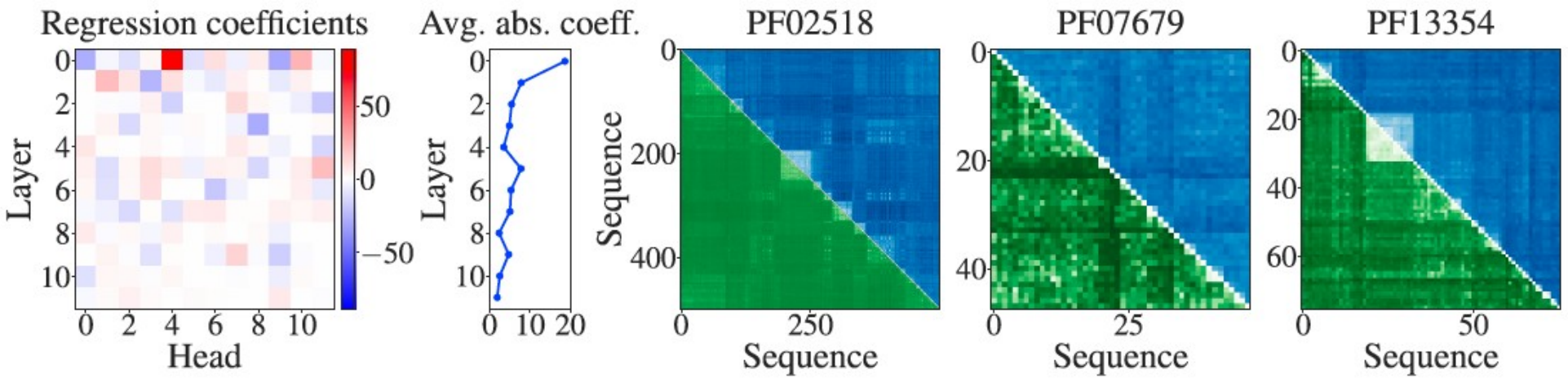
- **Global model**

  • We fit a global **logistic model** of the column attention matrices (averaged over columns) to predict the matrix of **pairwise Hamming distances** between sequences in MSAs – this time we train one model on multiple MSAs at once

  • Training: seed MSAs of 12 Pfam protein families

  • Test: seed MSAs of 3 other Pfam families

Hamming distance matrices and predictions for test MSAs

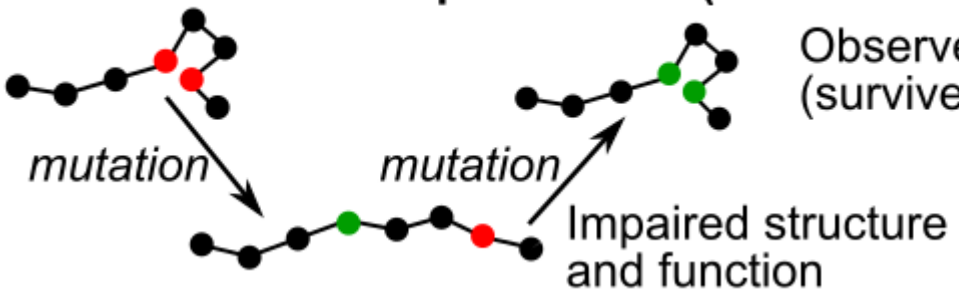


PF02518	PF07679	PF13354
R^2=0.60	R^2=0.28	R^2=0.67
ρ=0.90	ρ=0.85	ρ=0.92
Slope=1.2	Slope=0.57	Slope=0.70

Can MSA Transformer disentangle contacts and phylogeny?

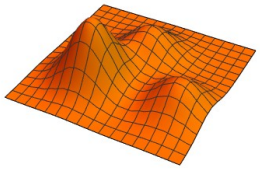
- **Correlations between columns can arise from contacts, but also from phylogeny**

Correlations from optimization (maintain structure and function):

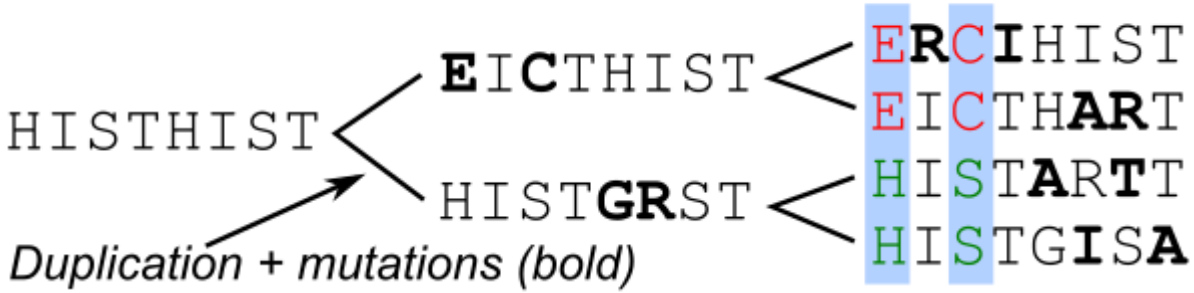


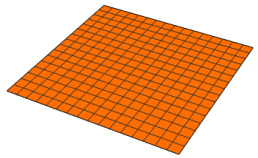
Observed sequences
(survived selection):

INTERACT
ANTGRANT
INAEPICT
INAGPANT



Correlations from historical contingency (phlogeny):





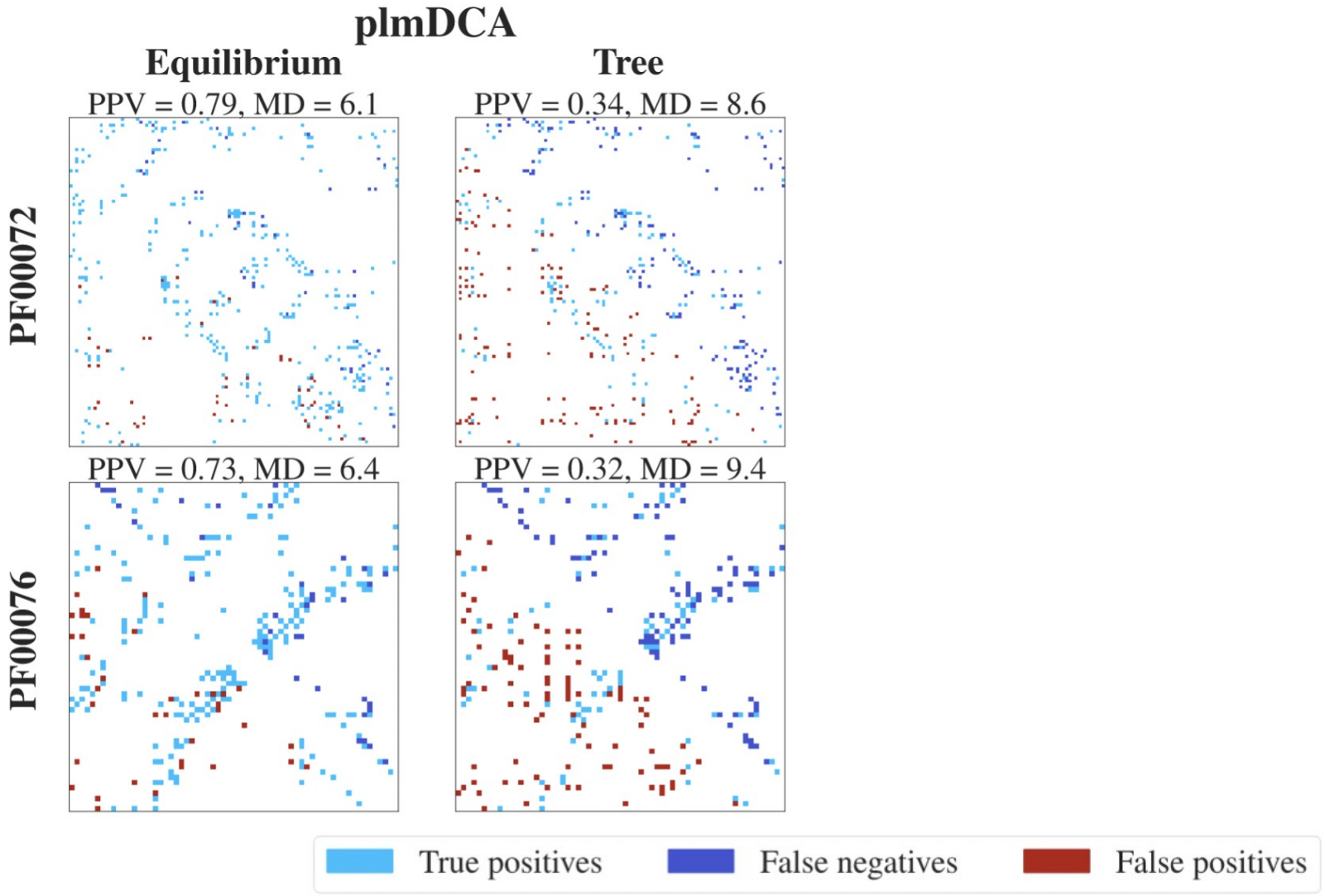
- **Correlations from phylogeny affect coevolution-based inference**

  • A problem for contact prediction by Potts models (Weigt, White et al 2009, Qin & Colwell 2018, Vorberg et al 2018, Hockenberry et al 2019, Malinverni & Barducci 2020, Rodriguez-Horta et al 2021, Colavin, Atolia et al 2022)

  • Useful for predicting interacting partners among paralogs (Marmier et al 2019, Gerardos et al 2021)

Can MSA Transformer disentangle contacts and phylogeny?

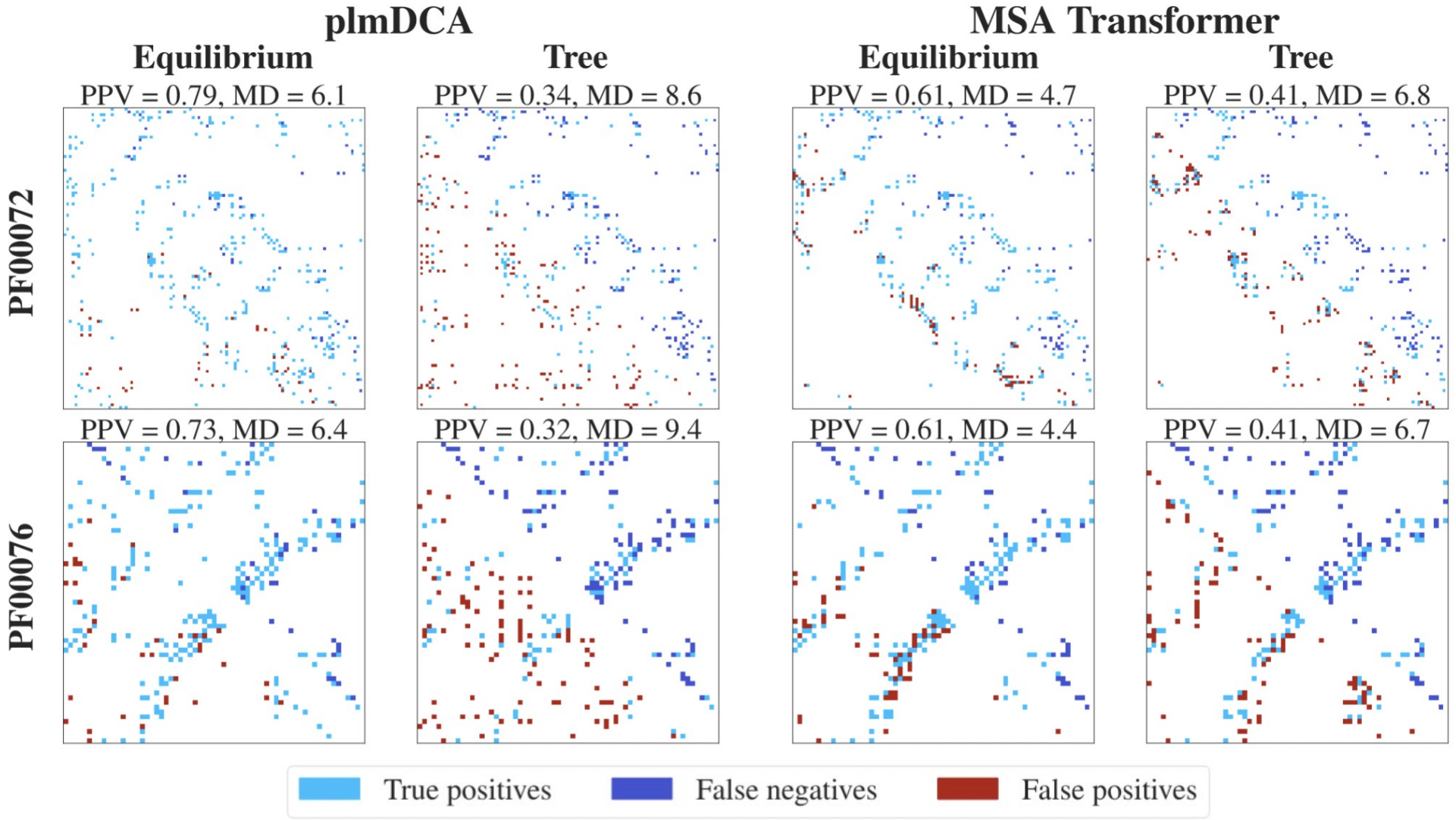
- **Robustness of contact prediction to phylogenetic noise (large MSAs – Pfam full)**



- Potts models (inferred by bmDCA on natural MSAs), sampled either:
  - Independently at equilibrium (Metropolis Hastings MCMC)
  - Along a phylogeny inferred from natural data (with the same Metropolis criterion)

Can MSA Transformer disentangle contacts and phylogeny?

- **Robustness of contact prediction to phylogenetic noise (large MSAs – Pfam full)**



- Potts models (inferred by bmDCA on natural MSAs), sampled either:
  - Independently at equilibrium (Metropolis Hastings MCMC)
  - Along a phylogeny inferred from natural data (with the same Metropolis criterion)
- MSA Transformer is more robust to phylogenetic noise than plmDCA

Conclusion

- **Summary**

  • MSA Transformer captures phylogenetic relationships within its column attentions

  • MSA Transformer has learnt phylogeny and contacts in orthogonal representations

  • MSA Transformer is more robust to phylogenetic noise than Potts models in the unsupervised contact prediction task

- **Perspective**

  • Can MSA-based language models be used to infer phylogenies?

- **Preprint**

  U. Lupo*, D. Sgarbossa and A.-F. Bitbol*, Protein language models trained on multiple sequence alignments learn phylogenetic relationships (2022) arXiv:2203.15465

Conclusion

- **Acknowledgments**

Joint work with:

Umberto Lupo (postdoc)
Damiano Sgarbossa (PhD student)

Discussions:

Mohammed AlQuraishi (Columbia)

CECAM workshop 2021 (organized by Alessandro Barducci, Duccio Malinverni, Paolo de los Rios)
EPFL POLS seminars
EPFL Ai4Science initiative

Current group members:

Alia Abbara (postdoc)
Celia García-Pareja (postdoc)
Umberto Lupo (postdoc)
Nicola Dietler (PhD student)
Richard Servajean (PhD student)
Damiano Sgarbossa (PhD student)
Alix Moawad (MSc student)
Mihaela-Diana Zanaoga (MSc student)
Evan Picchi (BSc student)

Current funding:

EPFL

erc
European Research Council
Established by the European Commission

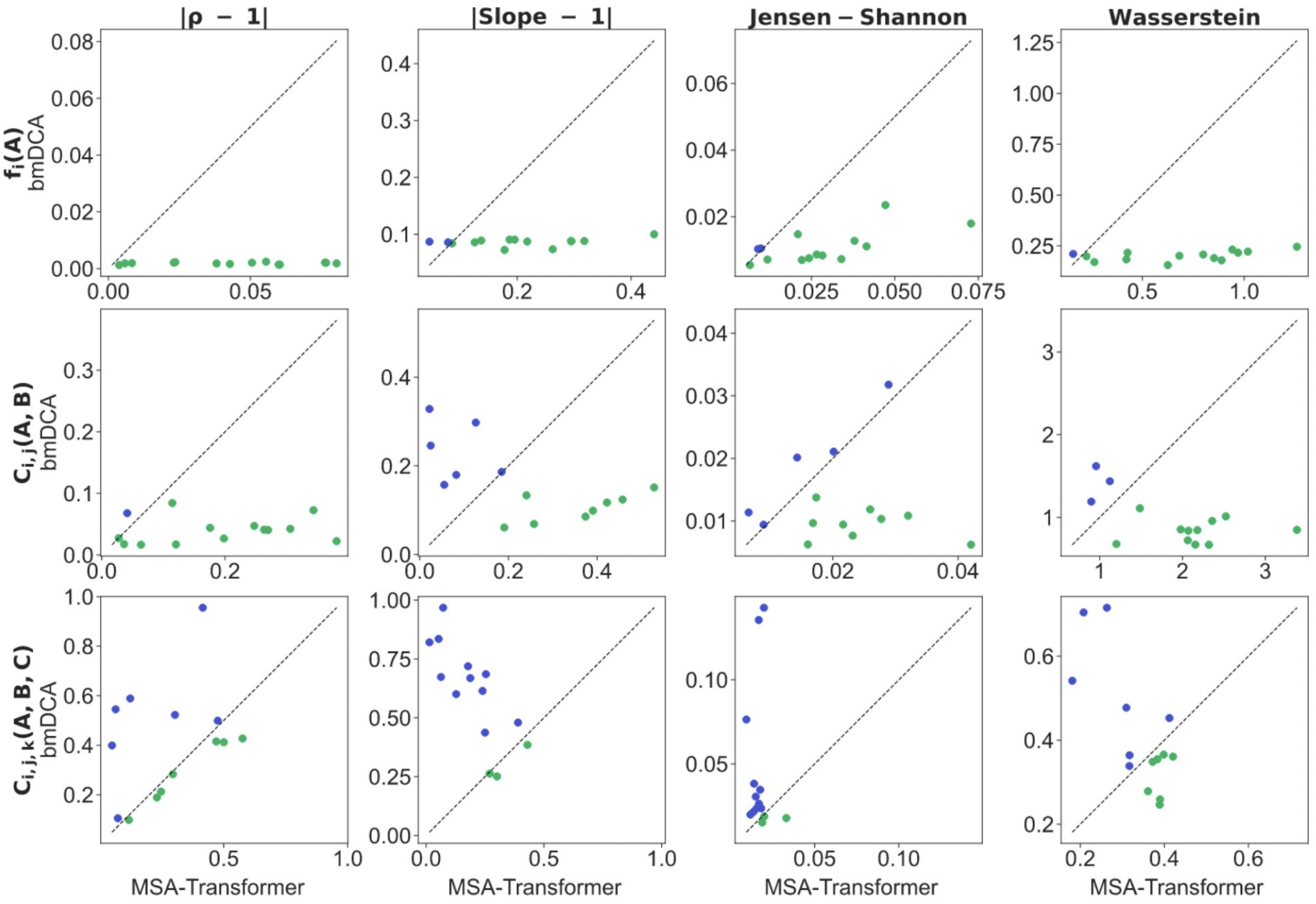
Group

Laboratory of Computational Biology and Theoretical Biophysics
Institute of Bioengineering, School of Life Sciences

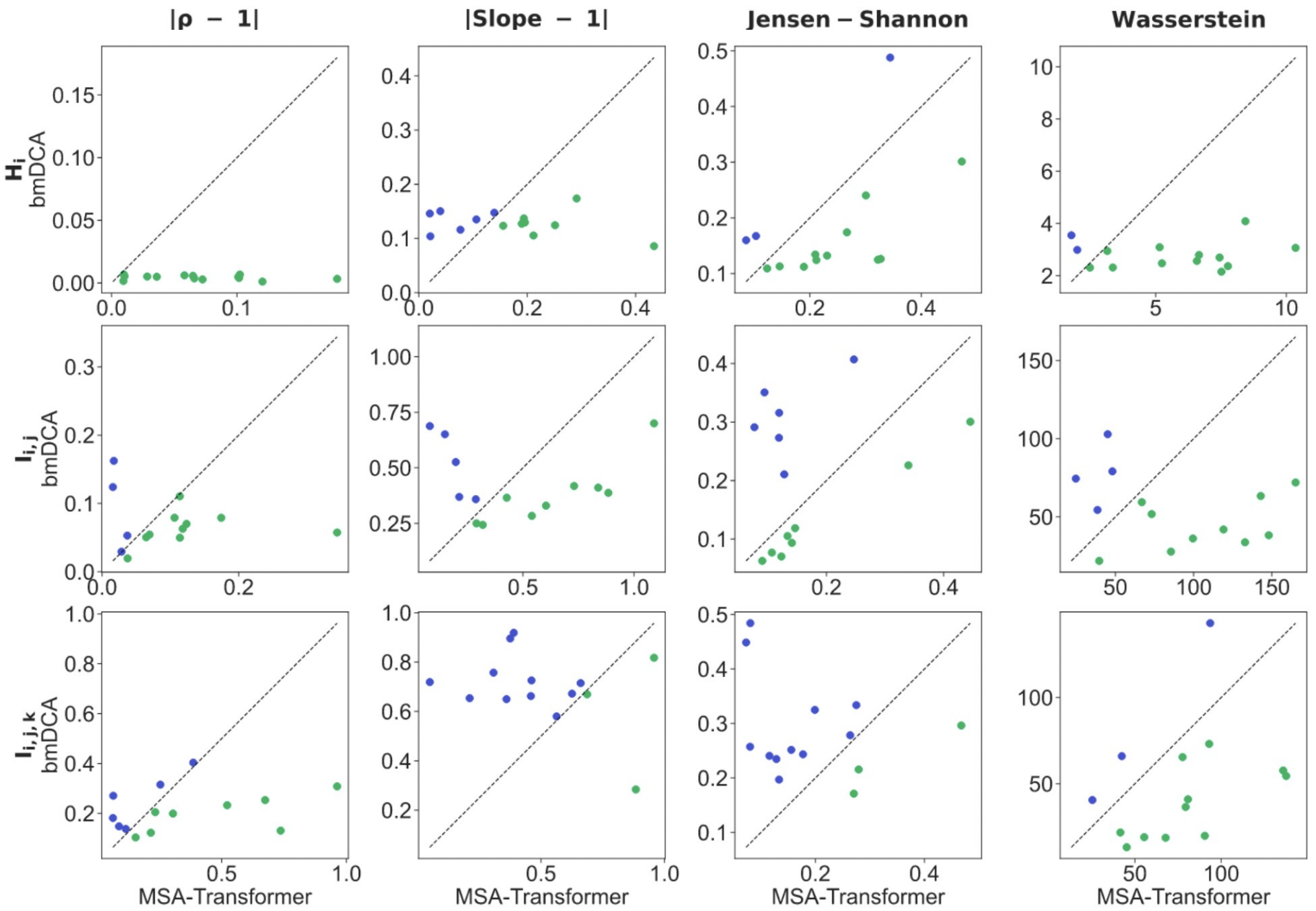
EPFL

Thanks!

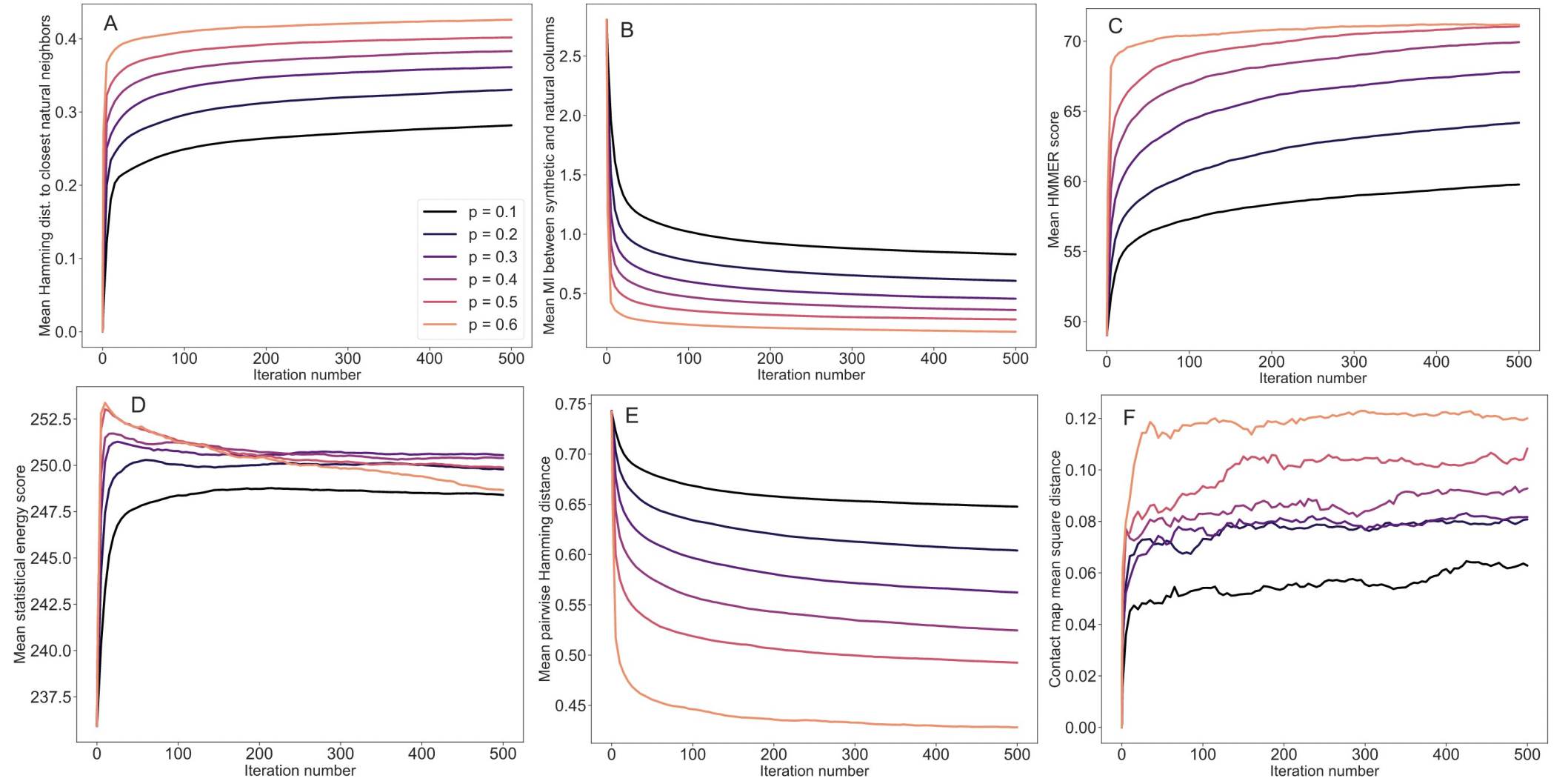
Comparison of statistics for all synthetic MSAs



Comparison of information measures for all synthetic MSAs



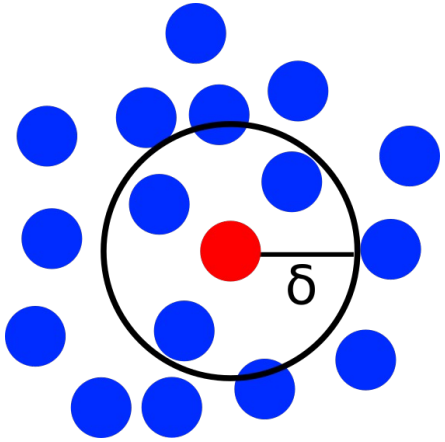
Impact of masking probability and number of iterations



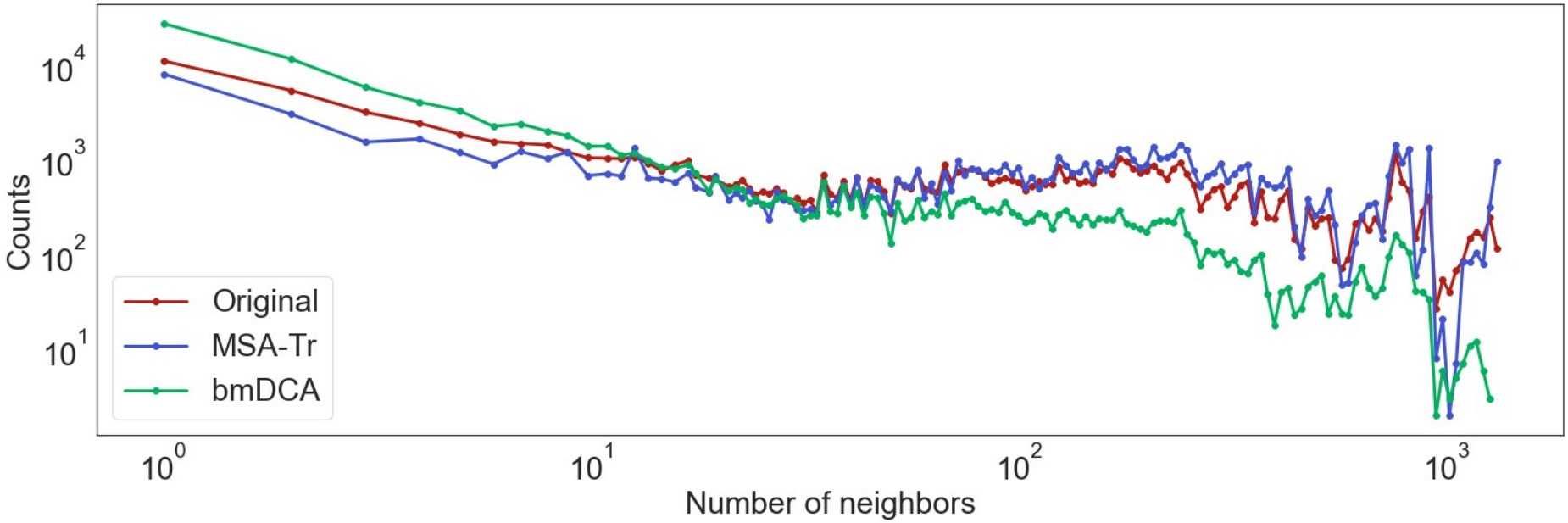
Impact of masking probability and number of iterations

Characterization of sequences generated by MSA Transformer

- **Number of neighbors of each sequence – PF00153 (mitochondrial ADP/ATP carrier)**



Number of neighbors = number of sequences (blue dots) inside a circle of radius δ (=0.2) centered on the sequence of interest (red dot)



Can MSA Transformer disentangle contacts and phylogeny?

- **Robustness of contact prediction to phylogenetic noise (large MSAs – Pfam full)**

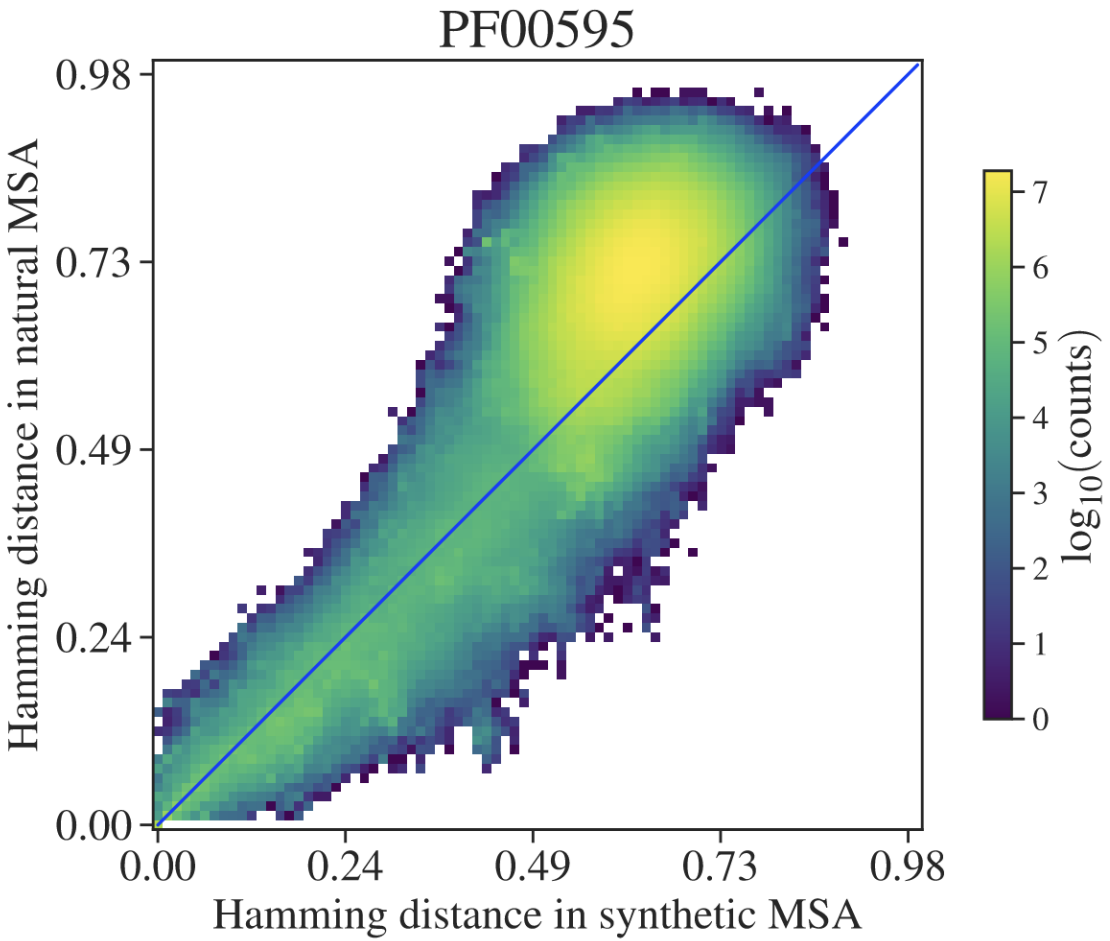
ROC AUC values; Δ = relative difference in %

	N contacts						$2L$ contacts					
	plmDCA			MSA Trans.			plmDCA			MSA Trans.		
Pfam ID	Eq.	Tree	Δ	Eq.	Tree	Δ	Eq.	Tree	Δ	Eq.	Tree	Δ
PF00004	0.87	0.58	33	0.70	0.67	4	0.93	0.61	34	0.80	0.71	11
PF00005	0.93	0.67	28	0.79	0.76	3	0.96	0.74	23	0.81	0.82	−1
PF00041	0.86	0.64	25	0.69	0.62	10	0.94	0.73	22	0.87	0.79	9
PF00072	0.94	0.73	23	0.86	0.77	10	0.99	0.85	14	0.94	0.87	8
PF00076	0.92	0.69	25	0.81	0.76	5	0.97	0.72	25	0.88	0.83	5
PF00096	0.88	0.54	39	0.68	0.54	21	0.92	0.54	41	0.78	0.54	30
PF00153	0.95	0.71	26	0.83	0.63	24	0.98	0.77	21	0.90	0.65	28
PF00271	0.91	0.62	32	0.78	0.72	7	0.95	0.67	29	0.85	0.77	10
PF00397	0.85	0.58	33	0.69	0.58	15	0.93	0.61	34	0.76	0.59	22
PF00512	0.94	0.74	21	0.84	0.77	8	0.97	0.78	20	0.88	0.81	8
PF00595	0.91	0.61	33	0.72	0.62	14	0.96	0.64	33	0.83	0.68	18
PF01535	0.85	0.66	23	0.66	0.63	05	0.88	0.72	18	0.73	0.72	1
PF02518	0.93	0.69	27	0.82	0.75	09	0.98	0.78	20	0.90	0.79	12
PF07679	0.85	0.63	26	0.68	0.64	05	0.95	0.77	19	0.85	0.80	5
PF13354	0.68	0.56	18	0.76	0.65	14	0.82	0.65	21	0.91	0.74	19
Average	0.88	0.64	27	0.75	0.68	10	0.94	0.71	25	0.85	0.74	12

Datasets

		Seed MSA		Full MSA			PDB structure	
Pfam ID	Family name	L	M	L	M	$M_{\text{eff}}^{(0.2)}$	ID	Resol.
PF00004	AAA	132	207	132	39277	9050	4D81	2.40 Å
PF00005	ABC_tran	137	55	137	68891	43882	1L7V	3.20 Å
PF00041	fn3	85	98	85	42721	17783	3UP1	2.15 Å
PF00072	Response_reg	112	52	112	73063	40180	3ILH	2.59 Å
PF00076	RRM_1	68	70	69	51964	20276	3NNH	2.75 Å
PF00096	zf-C2H2	23	159	23	38996	12581	4R2A	1.59 Å
PF00153	Mito_carr	97	160	94	93776	17860	1OCK	2.20 Å
PF00271	Helicase_C	111	421	111	66809	25018	3EX7	2.30 Å
PF00397	WW	31	448	31	39045	3361	4REX	1.60 Å
PF00512	HisKA	67	265	66	154998	67303	3DGE	2.80 Å
PF00595	PDZ	82	44	82	71303	4053	1BE9	1.82 Å
PF01535	PPR	31	458	31	109064	37514	4M57	2.86 Å
PF02518	HATPase_c	112	500	111	80714	59190	3G7E	2.20 Å
PF07679	I-set	90	48	90	36141	14611	1FHG	2.00 Å
PF13354	Beta-lactamase2	215	76	198	4642	3535	6QW8	1.10 Å

Generation of data along inferred phylogenies



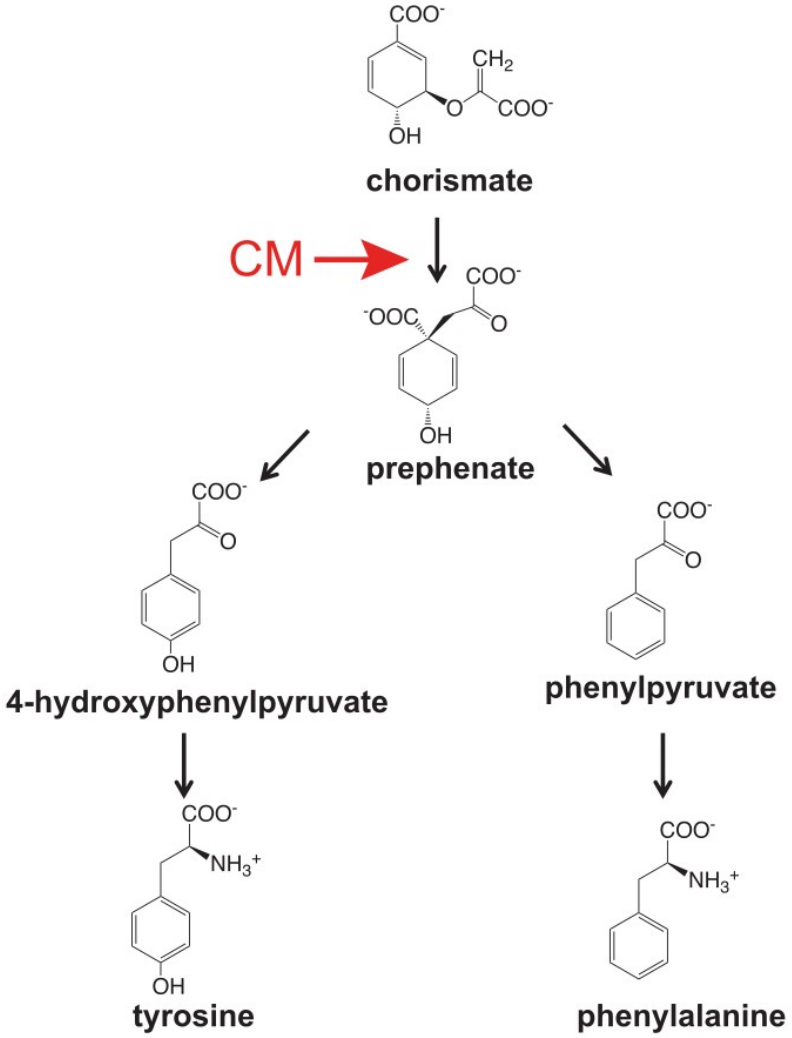
	Pearson correlation	
Family	All	$d_{\text{nat}} \leq 0.5$
PF00004	0.72	0.75
PF00005	0.31	0.62
PF00041	0.31	0.90
PF00072	0.33	0.85
PF00076	0.25	0.84
PF00096	0.32	0.23
PF00153	0.39	0.91
PF00271	0.37	0.70
PF00397	0.60	0.73
PF00512	0.30	0.65
PF00595	0.67	0.96
PF01535	0.15	0.70
PF02518	0.39	0.67
PF07679	0.19	0.90
PF13354	0.74	0.79

Beyond structure prediction: application to protein design

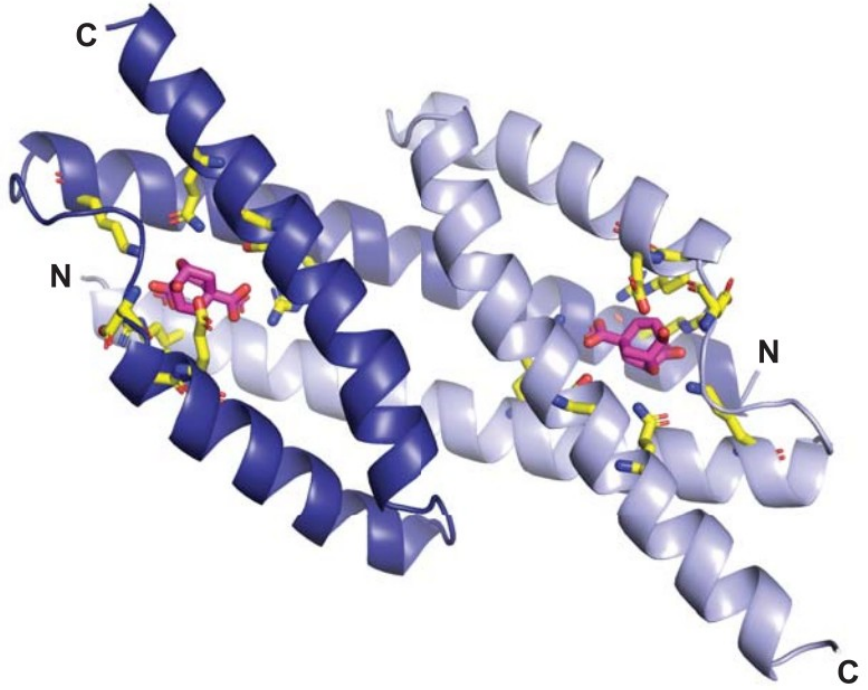
- **DCA (pairwise maximum entropy) models are generative**

  Russ et al, Science 369, 440–445 (2020)

Chorismate mutase (CM): a key enzyme in the biosynthesis of aromatic amino acids



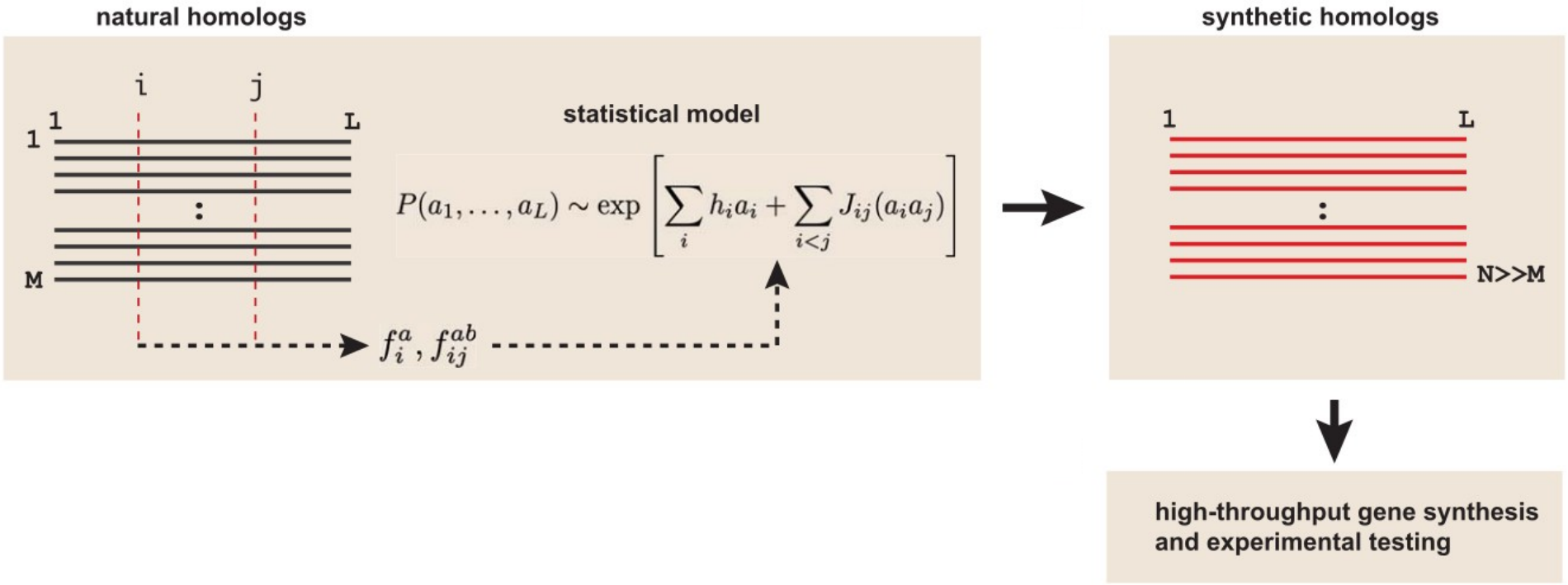
CMs form dimers. Active site residues are shown with yellow stick bonds and arise from both subunits. A bound substrate analog is shown in magenta.



Beyond structure prediction: application to protein design

- **DCA (pairwise maximum entropy) models are generative**
  Russ et al, Science 369, 440–445 (2020)

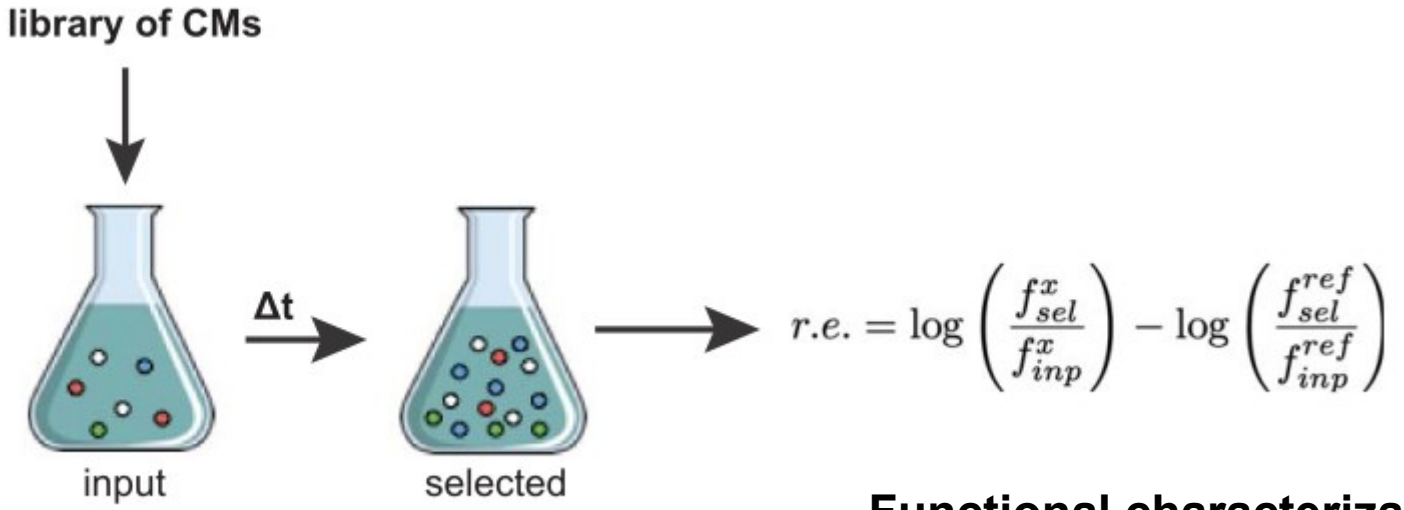
Using DCA to generate new CMs

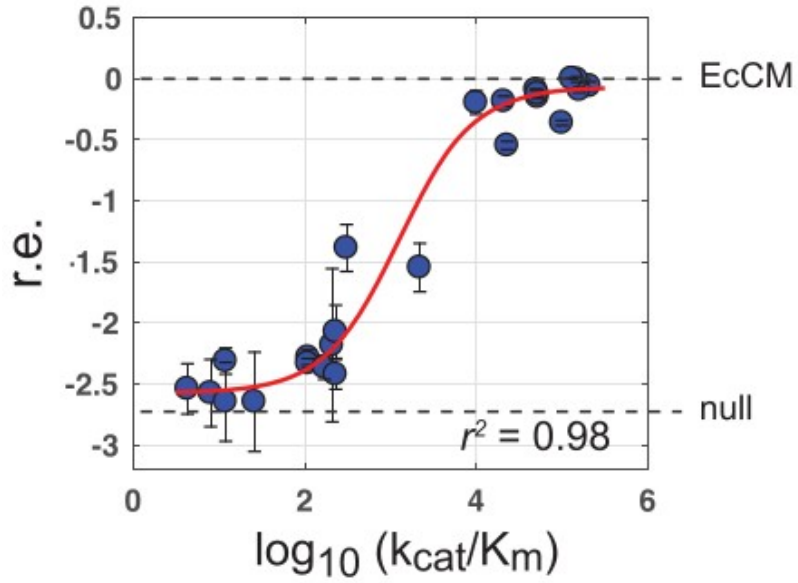


The model built on M natural CM homologs is used to generate N>>M artificial sequences that can be tested in a high- throughput assay for desired functions

Beyond structure prediction: application to protein design

- **DCA (pairwise maximum entropy) models are generative**
  Russ et al, Science 369, 440–445 (2020)





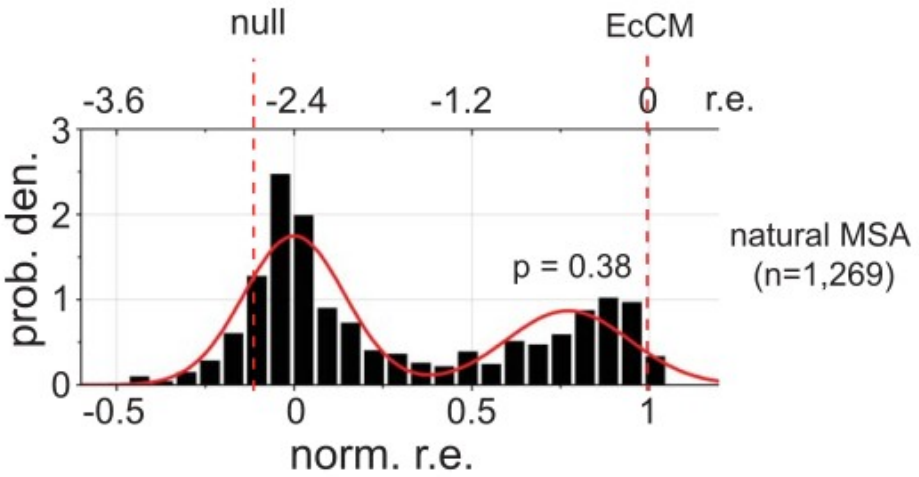
Functional characterization of CM activity:

- CM-deficient *E. coli* cells carrying libraries of variants are grown under selective conditions in minimal medium
- Deep sequencing of input and selected populations
- Calculation of the relative enrichment (r.e.) of each variant

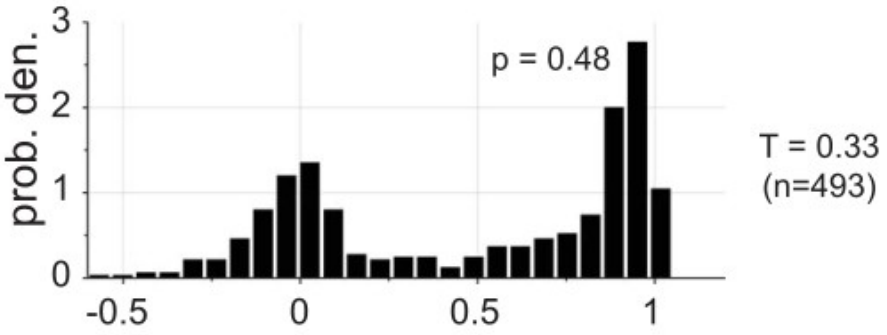
The relationship of r.e. to catalytic power [log 10 (k_{cat}/K_m)] for a number of CM variants yields a “standard curve” from complete lack of CM activity to wild-type EcCM activity

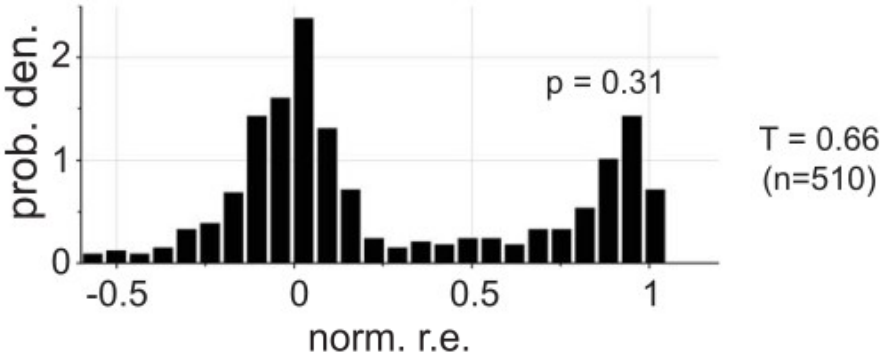
Beyond structure prediction: application to protein design

- **DCA (pairwise maximum entropy) models are generative**
  Russ et al, Science 369, 440–445 (2020)



The distribution of functional complementation by Natural AroQ sequences is bimodal, with ~38% of sequences in one mode near that of EcCM and the rest in another mode close to the r.e. of the null allele





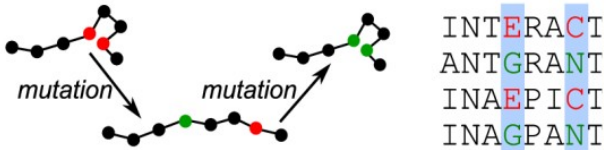
Distributions of r.e. values for artificial sequences sampled at T=0.33 and T=0.66 respectively

Overview of current projects in the group

Optimization & historical contingency

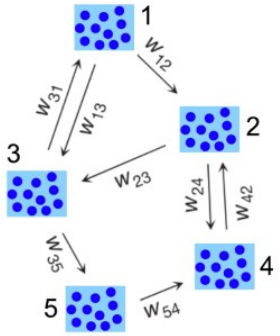
molecular scale

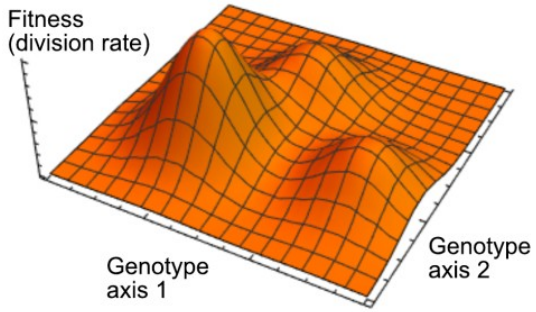
- **Axis 1:**
  Understanding how optimization & phylogeny shape protein sequences



population scale

- **Axis 2:**
  Assessing the predictability of evolution in spatially structured populations





Protein sequence data

- **Multiple sequence alignments**

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

Protein sequence data

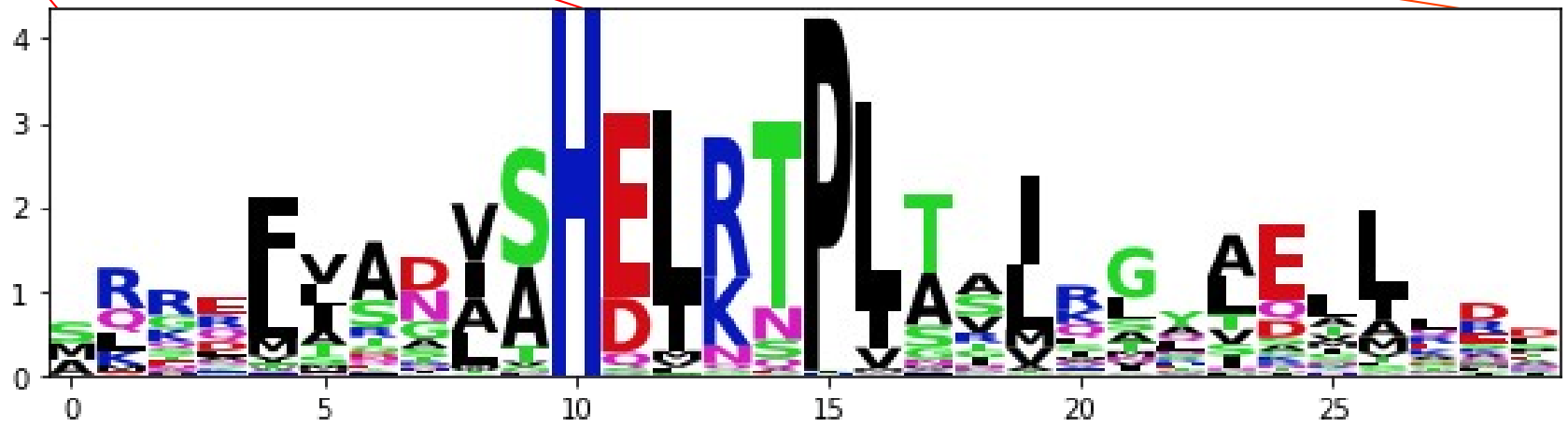
- **Multiple sequence alignments**

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

Protein sequence data

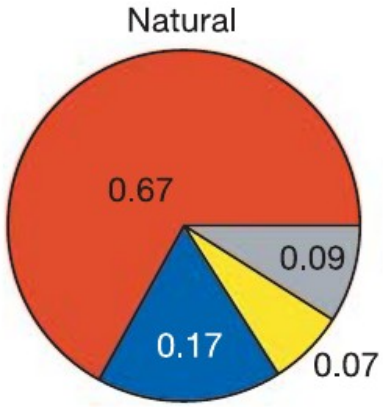
- **Multiple sequence alignments**

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



Correlations in sequences

- **(Conserved) correlations in amino acid usage are crucial**



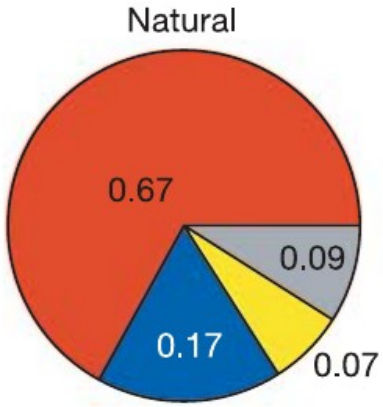
Socolich et al. 2005
synthetic WW domain

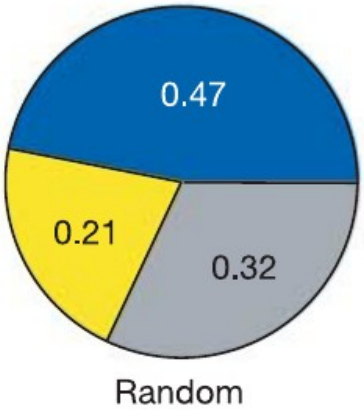
Red, natively folded
Blue, soluble but unfolded
Yellow, insoluble
Gray, poorly expressing

- Most natural sequences fold

Correlations in sequences

- **(Conserved) correlations in amino acid usage are crucial**





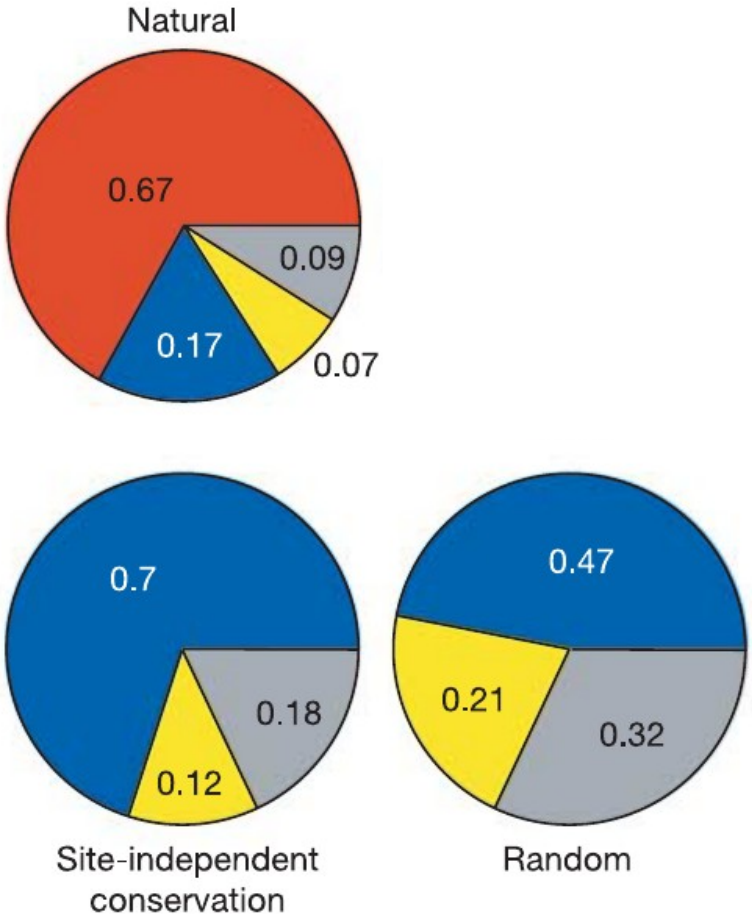
Socolich et al. 2005
synthetic WW domain

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

- Most natural sequences fold
- Random sequences don't

Correlations in sequences

- **(Conserved) correlations in amino acid usage are crucial**



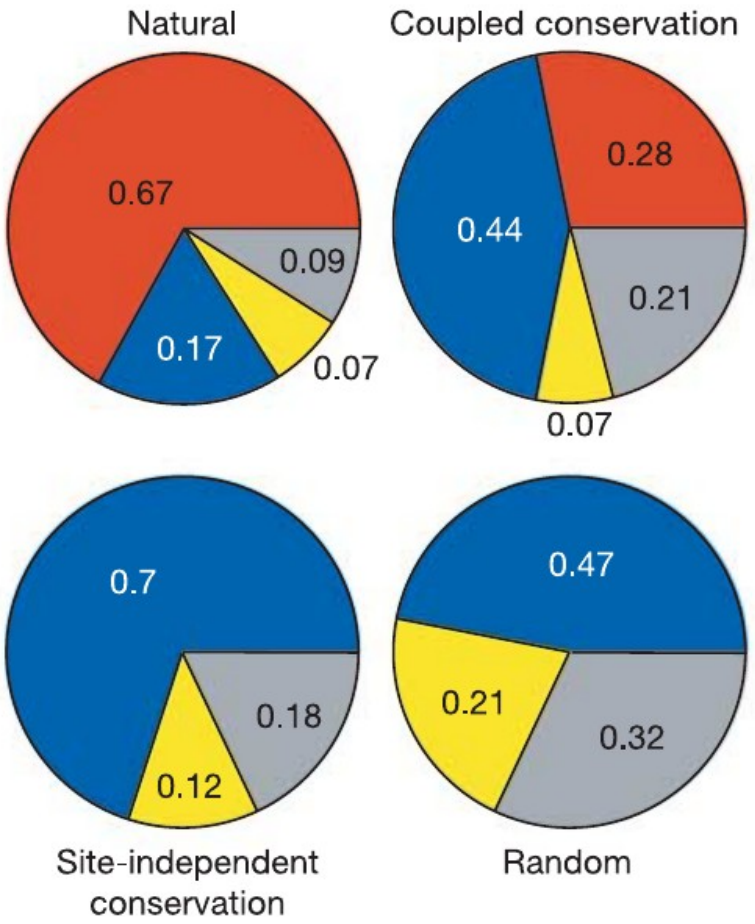
Socolich et al. 2005
synthetic WW domain

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 sequences

- **(Conserved) 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!

Deep learning and protein sequence data

- **Headlines**

AlphaFold: a solution to a 50-year-old grand challenge in biology Deepmind

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures Nature

A.I. Predicts the Shapes of Molecules to Come New York Times

AlphaFold Is The Most Important Achievement In AI—Ever Forbes

AlphaFold2 @ CASP14: "It feels like one's child has left home." M. AlQuraishi, Columbia

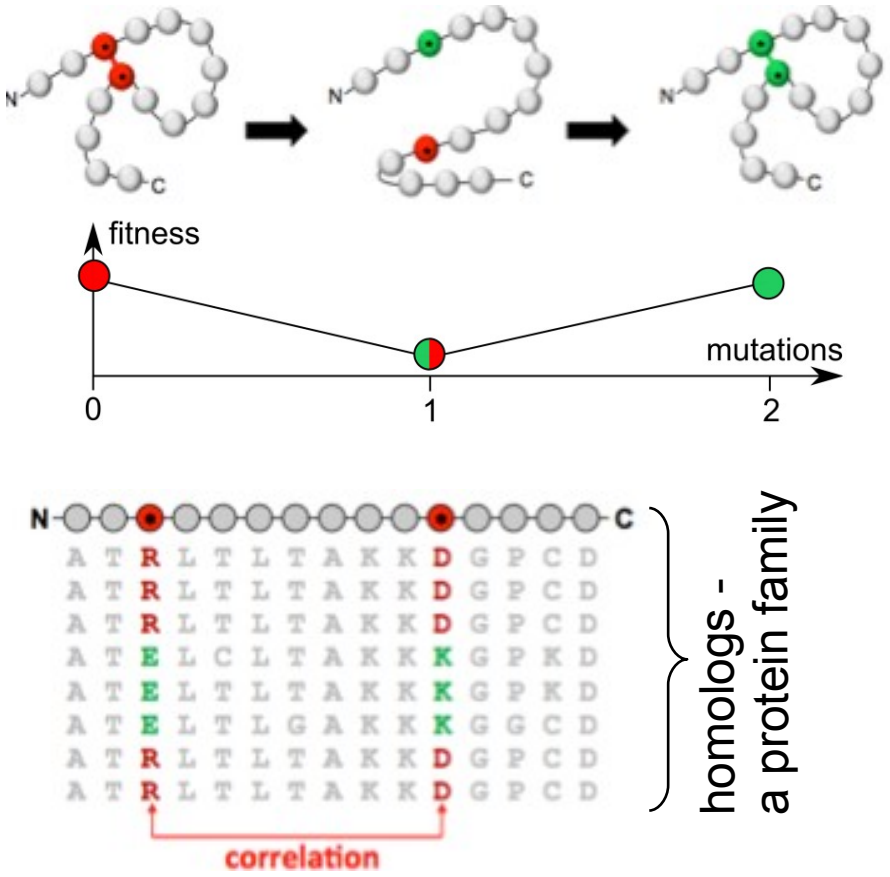
- **Other applications**

  - **Protein complex** structure prediction (Humphreys et al 2021)
  - **Protein-protein interactions** prediction (Evans et al 2021)

- **How about non-supervised learning using NLP-based architectures?**

Protein sequence data and inference

- **Inferring structure and function from sequences**



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

de Juan et al, 2013

- **Potts model (or Direct Coupling Analysis) –** Weigt, White et al 2009

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

Pairwise maximum entropy model

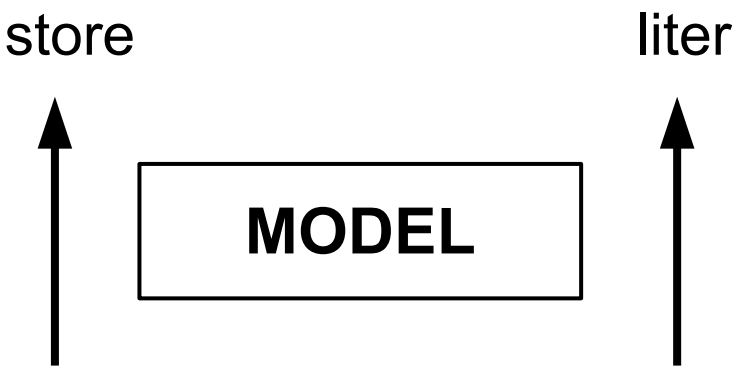
one-body terms - fields

two-body terms - (direct) couplings

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

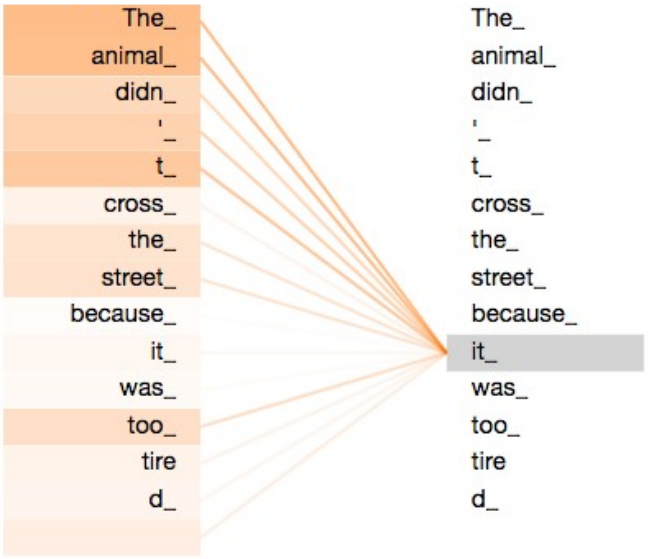
The model is trained to minimize a pseudo-likelihood loss:

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

Architecture of MSA Transformer

- **Transformer architecture**

One attention head



The Illustrated Transformer, Jay Alammar

M tokens → ***M* × *M* softmax** values

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 = 110M)

- **Adapting the transformer architecture to the 2D structure of protein MSAs –** Rao et al 2021

$BERT_{BASE}$-like model with **one-letter amino acids** as tokens, trained with MLM objective
Context for an amino acid is both its **row** and the **column** ("**axial attention**" – Ho et al 2019)

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