

Algorithms and Learning for Protein Science

From Darwin to AlphaFold:
MSA, DCA, attention mechanisms, and structure prediction

Frederic.Cazals@inria.fr

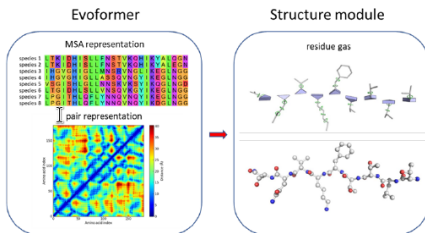
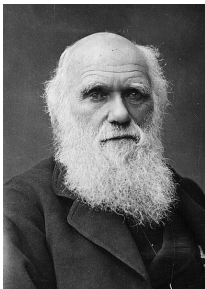


Figure 7. Schematic description of the two main modules of AF2. An input sequence together with data from sequence and structure databases serve as input to the Evoformer. The Structure module produces as output a 3D model of the protein structure corresponding to the input sequence.

Overview


▷ Theory/algorithms

- ▶ Potts models
- ▶ Transformers
- ▶ Topological persistence / Union-Find


▷ Protein science

- ▶ Analysis of Multiple Sequence Alignments (MSA)
- ▶ Contact predictions
- ▶ Generation of sequences
- ▶ AlphaFold and AlphaFold-DB


2024 Nobel prize in Chemistry



NOBELPRISET I KEMI 2024
THE NOBEL PRIZE IN CHEMISTRY 2024




KUNGL.
VETENSKAPS-
AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES




David Baker
University of Washington
USA

"för datorbaserad protein design"
"for computational protein design"



Demis Hassabis
Google DeepMind
United Kingdom

"för proteinstrukturprediktion"
"for protein structure prediction"



John M. Jumper
Google DeepMind
United Kingdom

- ▶ D. Baker: *For computational protein design*
- ▶ D. Hassabis and J. Jumper: *For protein structure prediction*

Perspective of the lecture

► AlphaFold:

- EvoFormer: supervised inference of contacts
- Structure module: production of the 3D structure

► This lecture:

- Unsupervised inference of contacts: slightly less accurate ... but explainable
 - Unsupervised: sequences only, no structures
- Assessment of AlphaFold reconstructions

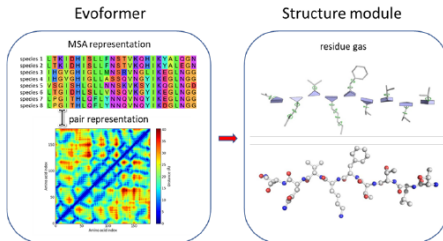


Figure 7. Schematic description of the two main modules of AF2. An input sequence together with data from sequence and structure databases serve as input to the Evoformer. The Structure module produces as output a 3D model of the protein structure corresponding to the input sequence.

From Darwin to AlphaFold

PART 1: MSA and DCA

PART 2: AlphaFold and AlphaFold-DB

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

Evolution, phylogeny, and multiple sequence alignments

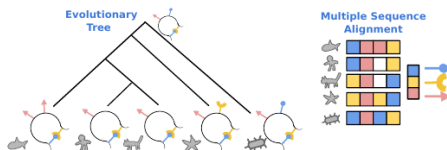


Figure 8: The tree on the right depicts evolution of a protein family. The protein at the root is the ancestral protein, and the five proteins at the leaves are its present-day descendants. The alignment on the left is the corresponding Multiple Sequence Alignment of observed sequences.



Figure 9: MSA for sequences from Figure 8 compared to a padded batch of the same sequences.

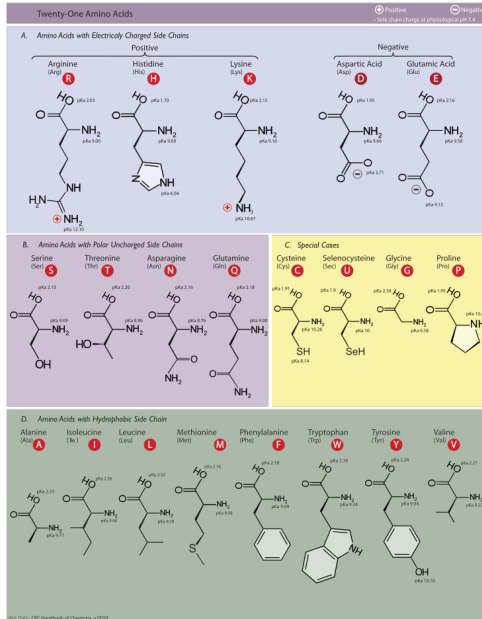
Figure: Evolution and multiple sequence alignments. From [1].

► **The PFAM database:** currently > 20,000 families and their MSA
<https://en.wikipedia.org/wiki/Pfam>,
<https://www.ebi.ac.uk/interpro/entry/pfam>

Multiple Sequence Alignments

Q5E940 BOVIN	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_HUMAN	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_MOUSE	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_RAT	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_CHICK	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_RANBY	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
Q7ZUG3 BRAHE	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_ICTPI	-----M	R	E	D	R	A	T	W	K	S	N	Y	F	L	K	I	L	D	D	Y	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	K	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	A	L	E	76									
RLA0_DROME	-----M	Y	R	E	N	K	A	W	A	Q	Y	I	K	V	E	L	D	E	F	T	K	C	F	I	V	G	A	D	N	V	G	K	K	M	Q	I	R	M	S	L	R	G	L	-	A	V	L	M	G	K	N	M	R	K	A	I	R	G	H	L	E	N	N	-	P	O	L	E	76									
RLA0_DICDI	-----M	S	G	A	G	-	S	K	R	K	N	V	F	E	K	A	T	K	L	F	T	T	D	K	M	I	V	A	E	A	D	F	V	G	S	L	O	L	K	I	R	K	S	I	R	G	I	-	G	A	V	L	M	G	K	N	M	I	R	K	V	I	R	D	L	A	D	S	K	-	P	E	L	D	75			
Q54LP0 DICDI	-----M	S	G	A	G	-	S	K	R	K	N	V	F	E	K	A	T	K	L	F	T	T	D	K	M	I	V	A	E	A	D	F	V	G	S	L	O	L	K	I	R	K	S	I	R	G	I	-	G	A	V	L	M	G	K	N	M	I	R	K	V	I	R	D	L	A	D	S	K	-	P	E	L	D	75			
RLA0_PLAF8	-----M	A	K	L	S	Q	Q	K	K	M	Y	T	E	L	S	S	L	I	Q	Q	S	K	I	L	I	V	H	V	D	N	V	G	K	N	M	A	S	V	R	K	S	L	R	G	K	-	A	T	L	M	G	K	N	I	R	T	A	L	K	N	L	-	P	O	L	E	76											
RLA0_SULAC	-----M	I	G	L	A	V	T	T	T	K	I	A	K	W	V	D	E	A	E	L	T	K	I	R	T	I	I	A	N	I	E	G	F	P	A	D	K	L	H	E	I	R	K	L	R	G	K	-	A	D	I	K	T	K	N	L	F	N	I	A	L	K	N	A	G	-	-	D	O	K	79							
RLA0_SULTO	-----M	R	I	M	A	V	I	T	Q	E	K	I	A	K	W	E	E	V	K	E	L	E	K	L	R	E	T	I	I	A	N	I	E	G	F	P	A	D	K	L	H	D	I	R	K	K	M	R	G	M	-	A	D	I	K	T	K	N	L	F	I	A	K	N	A	G	-	-	D	L	V	S	80					
RLA0_SULSO	-----M	K	E	L	A	L	A	L	Q	R	K	V	A	S	W	E	L	E	V	K	E	L	E	K	M	T	E	L	I	G	F	P	A	D	K	L	H	E	I	R	K	L	R	G	K	-	A	E	K	V	E	E	L	F	I	A	K	N	A	G	-	-	D	L	E	80												
RLA0_AERPE	-----M	G	W	V	E	L	G	Q	M	Y	K	R	E	K	T	E	W	K	T	L	M	L	E	E	L	F	S	K	W	V	V	L	F	A	D	I	T	S	E	E	V	V	E	R	V	H	K	K	W	K	-	P	O	M	V	A	K	K	I	L	E	A	M	K	A	A	L	E	-	-	L	D	M	86				
RLA0_PYRAR	-----M	H	M	L	A	I	G	R	R	Y	V	R	T	Q	P	A	R	V	K	I	V	S	E	A	T	E	L	I	K	Q	Y	V	F	L	D	L	H	G	L	S	R	I	L	H	E	V	Y	R	L	R	Y	-	G	V	I	K	I	K	P	L	F	K	I	A	F	T	K	V	Y	G	-	-	I	P	A	R	85	
RLA0_METAC	-----M	A	E	R	H	H	T	E	H	I	P	Q	W	K	K	D	E	T	E	N	K	E	L	Q	S	K	V	F	G	M	V	E	G	L	E	C	L	A	T	K	M	O	K	I	R	R	D	L	K	D	V	-	A	V	L	K	V	E	R	N	T	L	E	R	A	L	N	Q	L	-	-	E	T	I	D	78		
RLA0_METMA	-----M	A	E	R	H	H	T	E	H	I	P	Q	W	K	K	D	E	T	E	N	K	E	L	Q	S	K	V	F	G	M	V	R	I	E	G	L	A	T	K	I	O	K	I	R	R	D	L	K	D	V	-	A	V	L	K	V	E	R	N	T	L	E	R	A	L	N	Q	L	-	-	E	S	I	E	78			
RLA0_ARCFU	-----M	A	A	V	R	G	S	-	-	P	E	E	Y	K	V	R	A	V	E	E	K	R	M	I	S	S	K	V	V	A	I	V	S	E	R	N	V	P	A	G	O	M	O	K	I	R	R	E	F	R	G	-	A	E	K	V	V	K	N	T	L	E	R	A	L	D	A	L	G	-	-	D	O	V	L	78		
RLA0_METKA	-----M	A	V	A	K	K	E	P	O	P	S	G	E	Y	K	V	R	A	V	E	E	K	R	E	K	E	L	M	D	E	T	E	N	G	L	V	L	E	G	I	P	A	P	L	E	T	I	H	A	L	E	R	D	-	-	I	T	R	M	K	L	M	R	I	A	L	E	E	K	L	D	E	-	-	P	E	L	75
RLA0_METTH	-----M	A	H	V	A	E	W	K	K	E	V	E	E	L	A	N	L	I	K	S	Y	V	I	A	L	D	V	D	S	S	M	P	A	Y	L	S	Q	M	R	L	I	E	N	G	L	L	R	V	E	R	N	T	L	E	A	L	K	K	A	G	-	-	L	G	K	P	E	L	-	-	E	W	D	74				
RLA0_METTL	-----M	I	T	A	E	S	E	H	K	A	P	W	K	I	E	V	N	A	L	K	E	L	K	S	A	N	V	I	A	L	D	M	M	E	V	P	A	P	Q	L	E	I	R	D	K	I	R	-	E	T	L	K	M	K	M	R	N	T	L	E	R	A	E	V	A	E	T	G	N	P	E	A	-	-	P	E	A	82
RLA0_METVA	-----M	I	T	A	E	S	E	H	K	A	P	W	K	I	E	V	N	A	L	K	E	L	K	S	A	N	V	I	A	L	D	M	M	E	V	P	A	P	Q	L	E	I	R	D	K	I	R	-	D	O	M	L	K	M	R	N	T	L	E	R	A	E	V	A	E	T	G	N	P	E	A	-	-	P	E	A	82	
RLA0_METJA	-----M	E	T	K	V	K	H	V	A	P	W	K	I	E	V	N	A	L	K	E	L	K	S	K	P	V	A	I	D	M	M	E	V	P	A	P	Q	L	E	I	R	D	K	I	R	-	D	O	V	K	L	M	R	M	R	N	T	L	E	R	A	E	V	A	E	T	G	N	P	E	A	-	-	P	E	A	81	
RLA0_PYRAR	-----M	A	H	V	A	E	W	K	K	E	V	E	E	L	A	N	L	I	K	S	Y	V	I	A	L	D	V	D	S	S	M	P	A	Y	L	S	Q	M	R	L	I	E	N	G	L	L	R	V	E	R	N	T	L	E	A	L	K	K	A	G	-	-	L	G	K	P	E	L	-	-	E	W	D	77				
RLA0_PYRHO	-----M	A	H	V	A	E	W	K	K	E	V	E	E	L	A	N	L	I	K	S	Y	V	I	A	L	D	V	D	S	S	M	P	A	Y	L	S	Q	M	R	L	I	E	N	G	L	L	R	V	E	R	N	T	L	E	A	L	K	K	A	G	-	-	L	G	K	P	E	L	-	-	E	W	D	77				
RLA0_PYRFO	-----M	A	H	V	A	E	W	K	K	E	V	E	E	L	A	N	L	I	K	S	Y	V	I	A	L	D	V	D	S	S	M	P	A	Y	L	S	Q	M	R	L	I	E	N	G	L	L	R	V	E	R	N	T	L	E	A	L	K	K	A	G	-	-	L	G	K	P	E	L	-	-	E	W	D	77				
RLA0_PYRHO	-----M	A	H	V	A	E	W	K	K	E	V	E	E	L	A	N	L	I	K	S	Y	V	I	A	L	D	V	D	S	S	M	P	A	Y	L	S	Q	M	R	L	I	E	N	G	L	L	R	V	E	R	N	T	L	E	A	L	K	K	A	G	-	-	L	G	K	P	E	L	-	-	E	W	D	77				
RLA0_HALMA	-----M	S	E	S	E	R	K	T	E	T	E	W	K	E	V	D	A	I	V	M	I	E	S	T	S	V	G	Y	V	N	I	A	S	I	P	R	L	D	M	R	D	L	E	H	T	-	A	E	L	R	V	S	N	T	L	E	R	A	L	D	D	V	-	-	D	E	L	79										
RLA0_HALVO	-----M	S	E	S	E	R	V	T	E	V	P	W	K	R	E	V	D	E	V	D	F	I	E	S	T	S	V	G	Y	V	N	I	A	S	I	P	R	L	D	M	R	D	L	E	H	T	-	A	E	L	R	V	S	N	T	L	E	R	A	L	D	D	V	-	-	D	E	L	79									
RLA0_HALSA	-----M	S	E	S	E	R	V	T	E	V	P	W	K	R	E	V	D	E	V	D	F	I	E	S	T	S	V	G	Y	V	N	I	A	S	I	P	R	L	D	M	R	D	L	E	H	T	-	A	E	L	R	V	S	N	T	L	E	R	A	L	D	D	V	-	-	D	E											

The 20 natural amino acids (a.a.)



Speaking the language of proteins: Alphabets

Table 1. Amino acid alphabets and their clusters.^a

Alphabet	Clustering
UNIPROT20	A R N D C Q E G H I L K M F P S T W Y V
UNIPROT18	A R N D C Q E P G H L I K M F S T W Y V
HSDM17	A D K E R N T S Q Y F L I V M C W H G P
MMSEQS12	A S T L M I V K R E Q N D F Y C G H P W
WASS14	W M D I P C A V K T R E G L Y S H F N Q
SDM12	A D K E R N T S Q Y F L I V M C W H G P
GBMR7	D N A E F I K L M Q R V W Y C H T S G P
WWMJ5	C M F I L V W Y A T H G P D E S N Q R K
GBMR4	A D K E R N T S Q Y F L I V M C W H G P

^a Alphabet names are based on their literature abbreviations followed by a number denoting the number of clusters.

Figure: PLM and reduced a.a. alphabets. From

[2].

►Ref: Ieremie et al., PLM meet reduces a.a. alphabets, Bioinformatics, 2024

Contacts and their prediction

- ▷ **Example:** pairs of a.a. at distance $< 8\text{\AA}$ in 5ahw/chain A

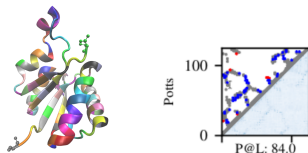


Figure: 5ahw-chainA: structure and contact prediction

- ▷ **Contacts assessment:** important parameters
- ▶ Sequence separation: number of residues in-between the two a.a. in contact: short: 6-11; medium: 12-23; long: 24+; extra-long: 50+ residues.
 - ▶ Precision at length: since the number of contacts grows linearly, prediction of the top 10, top L/5, top L/2, top L.
- ▷ **Usual statistics:** see <https://en.wikipedia.org/wiki/F-score>
- ▶ recall: percentage of positives $TP/(TP+FN)$
 - ▶ precision: fraction of true positives $TP/(TP+FP)$
 - ▶ F-score: $2 (precision \cdot recall) / (precision + recall)$

Number of pairwise contact: linear in protein length

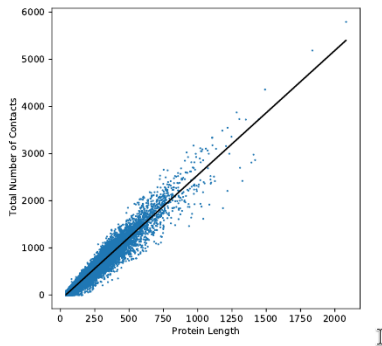


Figure: Number of contacts as a function of the protein length: linear. From [1].

▷ **Question:** implication for models ?

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

MSA and observables

► Assumptions:

- A sequence is denoted $A = (a_1, \dots, a_L)$, or $A^m = (a_1^m, \dots, a_L^m)$
- Consider a database \mathcal{D} of M protein sequences of length $< L$.
- A MSA has been computed: in matrix form

Matrix of shape $(M, L) : (a_i^m), i \in [L], m \in [M]$

- MSA alphabet with $q = 21$ symbols: 20 for the natural amino acids and the '-' for blanks in the MSA.
- All sequences: $\mathcal{A}^L - q^L$ of them

► **Frequency column-wise:** i -th column of the MSA, frequency of a.a. a is:

$$f_i(a) = \frac{1}{M} \sum_m \delta_{a, a_i^m}. \quad (1)$$

► **Frequency for two columns:** co-occurrence of a.a. a and b in columns i and j

$$f_{ij}(ab) = \frac{1}{M} \sum_m \delta_{a, a_i^m} \delta_{b, a_j^m}. \quad (2)$$

NB: adding up frequencies for pairs Eq. (2) yields frequencies for one column Eq. (1).

Correlations: pairwise and third-order

▷ **Rationale:** capture correlations between pairs/triples/etc of columns

Definition 1. The a.a. *pair correlation* is the $qL \times qL$ matrix whose entries are define by:

$$C_{ij}(a, b) = f_{ij}(ab) - f_i(a)f_j(b). \quad (3)$$

NB: in fact $(q-1)L \times (q-1)L$, see later

Definition 2. The third-order correlation tensor is defined by:

$$C_{ijk}(a, b, c) = f_{ijk}(abc) - f_{ij}(ab)f_k(c) - f_{ik}(ac)f_j(b) - f_{jk}(bc)f_i(a) + 2f_i(a)f_j(b)f_k(c). \quad (4)$$

▷ **Rmk.** The expression of the third order correlation stems from the expansion of

$$C_3(X, Y, Z) = \mathbb{E} [(X - \mu_X)(Y - \mu_Y)(Z - \mu_Z)]. \quad (5)$$

Modeling MSA: goals

- **Goal:** design a probability model for the sequences

$$\mathbb{P}[a_1, \dots, a_L], \quad (6)$$

so that it matches the observables from Eq. 1 and 2:

$$\mathbb{P}_i[a] = \sum_{A \in \mathcal{A}^L | a_i = a} \mathbb{P}[A] = f_i(a), \quad (7)$$

$$\mathbb{P}_{ij}[ab] = \sum_{A \in \mathcal{A}^L | a_i = a, a_j = b} \mathbb{P}[A] = f_{ij}(ab) \quad (8)$$

- **Two main applications:**

- Infer contacts – cf AlphaFold
- Sample non existing sequences – protein design

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

Direct Coupling Analysis (DCA): Potts models

▷ **Rationale:** define an *energy* model for the sequences, and apply Boltzmann's distribution

▷ **Hamiltonian:** parameters

▶ $h_i^{a_i}$: $L \times q$ parameters, one for each position and a.a. type.

▶ $J_{ij}^{a_i a_j}$: $\binom{L}{2} q^2$ parameters, one for each pair of positions and a.a. types.

With these, we define the following Hamiltonian of the sequence A :

$$H(A) = - \sum_i h_i^{a_i} - \sum_{i < j} J_{ij}^{a_i a_j} \quad (9)$$

NB: a **linear** model based on order 2 + order 4 tensors

▷ **Statistical model:** mimicking statistical physics

$$\mathbb{P}[A] = \frac{1}{Z} \exp(-H(A)), \text{ with } Z = \sum_{A \in \mathcal{A}^L} \exp(-H(A)). \quad (10)$$

▷ Ref: Weight et al, PNAS 2011, [3]

DCA model: free parameters

Lemma 3. The Hamiltonian model of Eq. 9 has $Lq + \binom{L}{2}q^2$ parameters, but $L + \binom{L}{2}(2q - 1)$ constraints.

Equivalently, the model has $L(q - 1) + \binom{L}{2}(q - 1)^2$ free parameters.

Proof. We have one constraint per position i , since

$$\forall i : \sum_a \mathbb{P}_i[a] = 1.$$

Consider now two positions i and j , and an a.a. a . We have

$$\forall i, \forall j, \forall a : \mathbb{P}_i[a] = \sum_b \mathbb{P}_{ij}[ab]$$

This yields $L(L - 1)q = \binom{L}{2}2q$ constraints. However, the constraints when $a = b$ have been counted twice, and we need to return $\binom{L}{2}$ degrees of freedom.

Therefore, the number of constraints is $\binom{L}{2}(2q - 1)$.

Equivalently, the model has $L(q - 1) + \binom{L}{2}(q - 1)^2$ free parameters. \square

The dependent parameters are set as follows [3]:

$$J_{ij}^{aq} = J_{ij}^{qa} = h_i^q = 0. \quad (11)$$

DCA model: applications

▷ Sequence generation:

- ▶ MCMC sampling with Metropolis-Hastings. Move set: changing one a.a. at a time.

▷ Contact prediction:

Definition 4. Using the order four tensor, the *contact strength* between positions i and j is given the following Frobenius norm:

$$\hat{C}_{ij} = \left\| J_{ij}^{\cdot\cdot} \right\|_F \quad (12)$$

- ▷ **Top contacts:** sort contacts by decreasing \hat{C}_{ij} , and compute the usual stats (precision, recall, F1, etc)

Properties and learning

- ▷ **Derivatives:** easily computed using the linearity in the Hamiltonian

Lemma 5.

$$\frac{\partial \log Z}{\partial h_i^a} = \mathbb{P}_i[a] \quad (13)$$

$$\frac{\partial^2 \log Z}{\partial h_i^a \partial h_j^a} = \mathbb{P}_{ij}[ab] - \mathbb{P}_i[a] \mathbb{P}_j[b]. \quad (14)$$

- ▷ **Learning:** based on pseudo-likelihood – omitted

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

Auto-regressive DCA model: arDCA

- ▷ **Rationale:** walk along the sequence
- ▷ **For a given i , the prefix in the sequence:**

$$(a_1, \dots, a_{i-1}) = \underline{a}_{i-1}. \quad (15)$$

With Bayes formula

$$\mathbb{P}[a_1, \dots, a_L] = \mathbb{P}[a_1] \mathbb{P}[a_2 | a_1] \dots \mathbb{P}[a_L | \underline{a}_{L-1}] \quad (16)$$

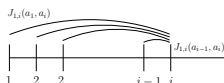
The auto-regressive term aims at predicting the symbol a_i . We compute it as follows:

$$\mathbb{P}[a_i | \underline{a}_{i-1}] = \frac{\exp(h_i^{a_i} + \sum_{j=1, \dots, i-1} J_{ij}^{a_i a_j})}{z_i(a_{i-1}^a, \dots, a_1^a)}, \quad (17)$$

with

$$z_i(a_{i-1}^a, \dots, a_1^a) = \sum_{a_i} \exp(h_i^{a_i} + \sum_{j=1, \dots, i-1} J_{ij}^{a_i a_j}) \quad (18)$$

NB: the expression involves $1 + i - 1$ parameters which are specific to position i , namely $h_i^{a_i}$ and $J_{ij}^{a_i a_j}$.



The prediction of a_i involves
 $1 + i - 1 = i$ parameters. From [4].
See also [5]

Figure: Auto-regressive model to predict a_i .

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

Attention and weighted sums

Consider a $n \times d$ matrix $P = [v_1, \dots, v_n]^T$ of n vectors v_i in \mathbb{R}^d .

The product of P and its transpose yields the covariance matrix and the Gram matrix:

$$\begin{cases} C = P^T P = (C_{ij})\text{-shape } (d, d) \\ G = P P^T = (\langle v_i, v_j \rangle)\text{-shape } (n, n). \end{cases} \quad (19)$$

Consider the following transformation, which takes a weighted sum of all vectors, weighted by the soft max of the dot products:

$$v'_i = \sum_j \frac{\exp(-\langle v_i, v_j \rangle) v_j}{\sum_j \exp(-\langle v_i, v_j \rangle)}. \quad (20)$$

For the n vectors, this transformation reads as follows in matrix form:

$$P' = \text{softmax}_r(-P P^T) P. \quad (21)$$

This is the weighted sum implemented by transformers and attention mechanisms. Note that here, the soft max is taken row-wise, whence the subscript.

Attention: single head

- ▷ **Attention:** with data are vectors of shape $(1, d_m)$, weights matrices W_Q, W_K, W_V

	Data	Projection matrix	Result
Queries	$X : n \times d_m$	$W_Q : d_m \times d_k$	$Q = XW_Q : n \times d_k$
Keys		$W_K : d_m \times d_k$	$K = XW_K : n \times d_k$
Values		$W_V : d_m \times d_v$	$V = XW_V : n \times d_v$

- ▷ **Single head attention.** The attention mechanisms consider three projection matrices (to be learned), yielding the so-called Query, Key, and Value matrices:

$$\text{Query: } Q = XW_Q, \text{ Key: } K = XW_K, \text{ Value: } V = XW_V. \quad (22)$$

- ▷ **Attention score matrix:** the $n \times n$ matrix coding the *attention* each token has for every other token:

$$QK^T. \quad (23)$$

- ▷ **Sparsifying and rescaling with softmax_r :** yields the final embedding *i.e.* a matrix of shape $n \times d_v$

$$\text{softmax}_r\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (24)$$

NB: dimension-wise: $(n \times d_k)(d_k \times n)(n \times d_v) = (n \times d_v)$.

Q vs W_Q , etc: example

- ▷ Consider a MSA: size (M, L)
- ▷ Position i of the sequence x : two types of attributes:
 - ▶ sequence info: the a.a. type,
 - ▶ position info: some positional information, e.g. the secondary structure type, or a vector encoding biophysical properties.

Assume that both encodings are represented as vector of \mathbb{R}^{d_m} , we obtain:

$$x'_i = E_{\text{seq}} + E_{\text{pos}}.$$

Then, the matrix Q is obtained as $Q = X'W_Q$, with $X' = x'_1, \dots, x'_m{}^T$.

▷Ref: Example from [1]

Attention: multiple heads

- ▶ Using H attention heads: with dimensions $d_k = d_v = d_m/h$
 - ▶ Each individual head computes

$$\text{softmax}_r({}^h Q {}^h K^{\text{T}}) {}^h V. \quad (25)$$

- ▶ Concatenating the H outputs yields a matrix of shape $n \times d_m$.

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

Factored attention: rationale

▷ **DCA limitations:** model size $L(q-1) + \binom{L}{2}(q-1)^2$ is suboptimal in two respects:

- ▶ Two pairs of a.a. with identical properties yield two terms which will be alike
Recall that charged interactions correspond to: $\{ (R) \text{ arginine}, (K) \text{ lysine}, (H) \text{ histidine} \} \times \{ (D) \text{ aspartic acid}, (E) \text{ glutamic acid} \}$.

Therefore, we expect:

$$J_{ij}^{DR} \equiv J_{ij}^{EK}$$

- ▶ In a protein, the number of contacts is linear and not quadratic in the protein length [6].

Problem 6. Define a Hamiltonian subquadratic in size, ideally linear, exploiting universal properties of a.a. types.

Factored attention: model

- ▷ **Simplification 1 wrt DCA:** focus on pairwise interactions, discard the terms $h_i^{a_i}$
- ▷ **With head-size d (hyper-parameter):** consider
 - ▶ W_Q : $L \times d$ matrix, position dependent,
 - ▶ W_K : $L \times d$ matrix, position dependent,
 - ▶ W_V : $q \times q$ matrix, depending on a.a. properties.
- ▷ **Coupling term:** expression

$$J_{ij}^{a_i a_j} = \text{symmetrized}(\text{softmax}_r(W_Q W_K^T))_{ij} W_V(a_i, a_j). \quad (26)$$

- ▷ **Comments are in order:**
 - ▶ The previous eq. decouples positions and a.a. properties
 - ▶ The $L \times L$ matrix $W_Q W_K^T$: self attention for pairs of a.a.
 - ▶ Model size: $2dL + q^2$ instead of $O(L^2 q^2)$ for a classical DCA model.
 - ▶ The matrix W_V encodes pairwise interactions between a.a. types.
- ▷ **With multiple heads:**

$$J_{ij}^{a_i a_j} = \sum_h \text{symmetrized}(\text{softmax}_r(W_{Q_h} W_{K_h}^T))_{ij} W_{V_h}[a_i, a_j]. \quad (27)$$

Factored attention: model sizes

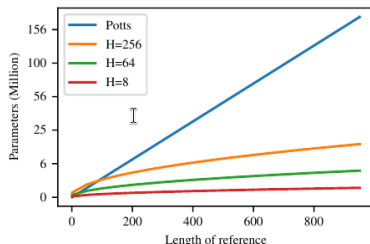


Figure 18: Number of parameters versus length for MRF models.

Figure: fig18 **Models' sizes: Pott's models versus transformers, for head size $d = 18$.** Potts requires a total of 12 billion parameters to model all 748 families. Factored attention with 256 heads and head size 32 has 3.2 billion parameters; lowering to 128 heads reduces this to 790 million. Half of this reduction comes from 107 families of length greater than 400. ProtBERT-BFD is the most efficient, with 420 million parameters. From [1].

Factored attention: results (I)

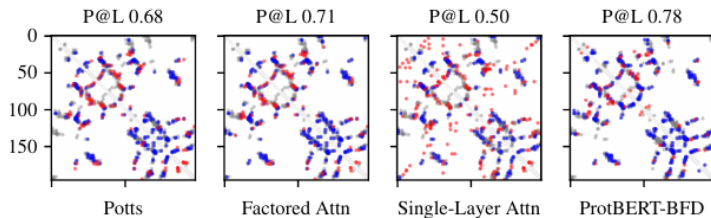


Figure 2: Predicted contact maps and Precision at L for each model on PDB entry *2BFW*. Blue indicates a true positive, red indicates a false positive, and grey indicates a false negative.

Factored attention: results (II)

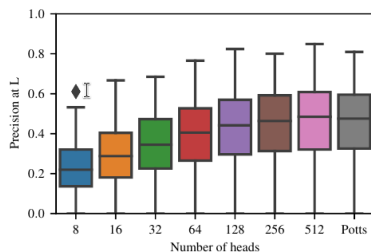
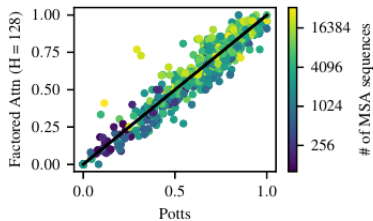
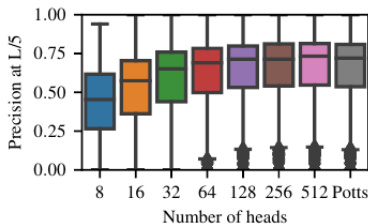


Figure: fig13 Performance as a function of the number of heads.
From [1].

Factored attention: results (III)



(a) 128 heads versus Potts $P@L/5$



(b) $P@L/5$ for varying number of heads.

Figure: fig5 **Incidence of the number of heads on precision. (Left)** Impact of the number of heads on precision at $L/5$. NB: one point per MSA/protein family. **(Right)** Sharp decline with fewer than 32 heads. From [1].

From Darwin to AlphaFold

Multiple sequence alignments and protein contacts

MSA: observables and models

Direct Coupling Analysis

The auto-regressive DCA model: arDCA

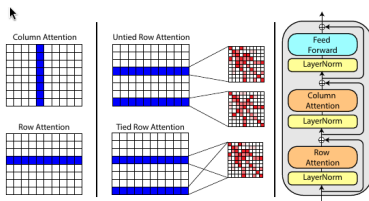
Attention and transformers

Coupling analysis with factored attention

Coupling analysis with tied attention: MSA transformer

MSA transformer: rationale

- ▶ **Axial attention for a MSA of size $M \times L$:**
 - ▶ Row based: yields size $O(ML^2)$
 - ▶ Column based: yields size $O(LM^2)$
- ▶ **MSA transformer:**
 - ▶ aggregate information across sequences / lines of the MSA
 - ▶ constrains the contact structure



- ▶ **Tied attention for rows:**

$$\sum_{m=1}^M \frac{Q_m K_m^T}{\lambda(M, d)}, \quad (28)$$

NB: MSA transformer block

- ▶ Row then column attention

Figure: Axial attention versus tied attention. From [7].

- ▶ Model size: $O(L^2)$.
- ▶ The std \sqrt{d} normalization is replaced by:
 - $\lambda(M, d) = M\sqrt{d}$ or $\lambda(M, d) = \sqrt{Md}$ – used in practice.

MSA transformer: results (I)

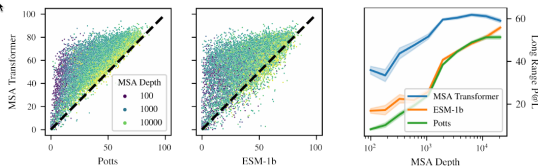


Figure: fig2 Top-L long-range contact prediction – higher is better.
The MSA transformer performs well with low depth. From [7].

Table 2. Unsupervised contact prediction on CASP13 and CAMEO (long-range precision). Note the large improvement of MSA Transformer over classical Potts models and ESM-1b.

Model	CASP13-FM		CAMEO	
	L	L/5	L	L/5
Potts	16.9	31.5	24.0	42.8
ProTrans-T5	16.5	27.0	25.9	43.4
ESM-1b	17.0	30.4	30.9	52.7
MSA Transformer	43.4	71.1	43.4	66.2

Table 3. Supervised contact prediction on CASP13 and CAMEO (long-range precision). *Uses outer-concatenation of the query sequence representation as features. †Additionally uses the row attention maps as features.

Model	CASP13-FM		CAMEO	
	L	L/5	L	L/5
trRosetta _{base}	45.7	69.6	50.9	75.5
trRosetta _{full}	51.8	80.1	53.2	77.5
Co-evolutionary	40.1	65.2	47.3	72.1
ProTrans-T5	25.0	41.4	40.8	63.3
ESM-1b	28.2	50.2	44.4	68.4
MSA Transformer*	54.5	80.2	53.6	78.0
MSA Transformer†	54.6	77.5	55.8	79.1

Table: Unsupervised prediction of contacts.

MSA transformer: results (II)

- ▶ Pott model: uses the covariance information in the MSA
- ▶ PLM: uses patterns in the sequence
- ▷ **Braking both inference models:**
 - ▶ Covariance ablation: random permutation within a MSA column; preserves the a.a. frequencies, but brakes correlations.
 - ▶ Sequence patterns ablation: permute columns in the MSA; covariance information between pairs of columns preserved, but *scrambled* sequences.

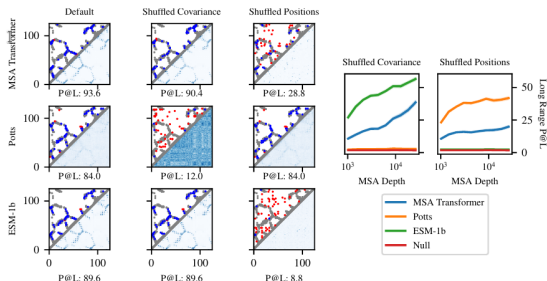


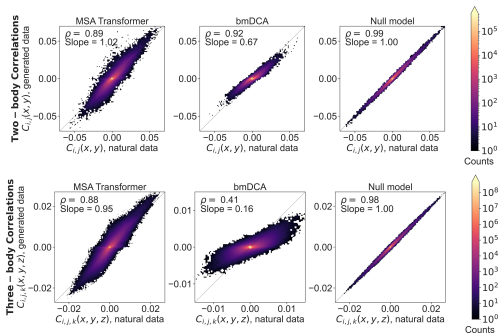
Figure: fig6 Learning modes: covariance of sequence patterns Left: one example. Right: results binned per MSA depth. Potts breaks with the covariance ablation; EMS break with sequence position ablation. From [7].

MSA transformer vs DCA models: third order correlations

► Protocol:

- Sample sequences using MSA transformer and DCA
- Compare statistics against those of PFAM families

► NB: higher order correlations better captured



►Ref: Sgarbossa et al, eLife 2023; [8]

From Darwin to AlphaFold

PART 1: MSA and DCA

PART 2: AlphaFold and AlphaFold-DB

From Darwin to AlphaFold

Intermezzo: Union-Find

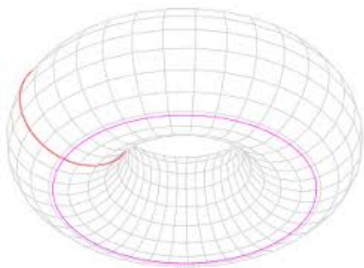
AlphaFold predictions at a glance

Packing analysis: methods

pLDDT: methods

Results

Topological invariants: Betti numbers



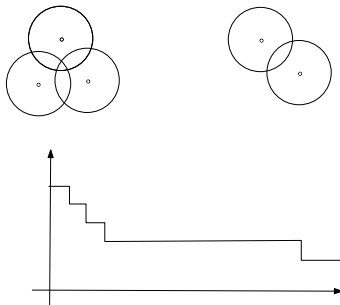
▷ **Torus / doughnut / inner tube:** $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$

▷ **Euler characteristic:**

$$\chi = \sum_i (-1)^i \beta_i. \quad (29)$$

What we see/like is stable – at certain scale

▷ **Clustering:** how many clusters ?



A plausible number of clusters corresponds to a plateau \Rightarrow connected components maintained with the Union-Find algorithm

The Union-Find problem

► Connected components of a graph:

- Static graph: run a depth first search algorithm
- Dynamic graph: Union-Find

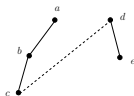
Problem 7. Consider a dynamics graph into which nodes and edges are being added. Problem: maintain the connected components over time.

The principles behind Union-Find are pretty simple (Fig. ??):

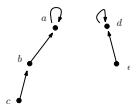
1. Represent a c.c. by a rooted tree: the leader of the c.c. is the node of the tree.
2. Find : find the leader of a node
3. Union : union 2 components

► Union-Find example:

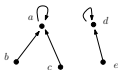
Graph



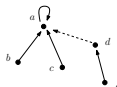
Its representations as a forest of trees



Forest after the Find operations



Forest after the Union operation



Connecting two nodes i and j yields the merge of their connected components. The c.c. are represented by trees, whose structure changes over the operations.

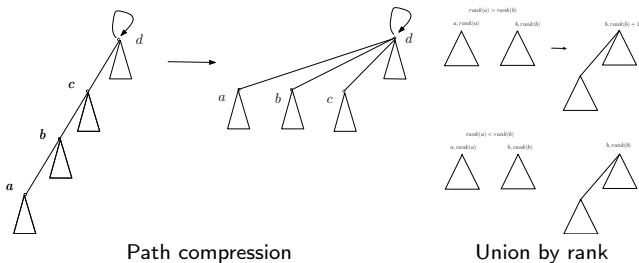
Union-Find: heuristics

▷ UF.find uses path compression:

- ▶ Action: find uses path compression: any node on the path from the query to the root (leader) is attached to the root.
- ▶ (Rationale) Path compression useful for future queries – *amortized* analysis.

▷ UF.union uses Union-by-rank:

- ▶ Action: the tree with largest rank remains the leader.
- ▶ Rationale: keep trees shallow.



Union-Find: complexity

Definition 8. The *Ackermann* function is defined by:

$$A(1, j) = 2^j, j \geq 1$$

$$A(i, 1) = A(i - 1, 2), i \geq 2$$

$$A(i, j) = A(i - 1, A(i, j - 1)), i, j \geq 2$$

Its inverse is defined by

$$\alpha(m, n) = \min\{i \geq 1 : A(i, \lfloor m/n \rfloor) > \log n\}$$

- ▶ For n fixed, $\alpha(m, n)$ is decreasing in m
- ▶ $\alpha(m, n)$ is ≤ 5 for all practical purposes

Theorem 9. A sequence of m union-find operations on a n elements set has complexity $O(m\alpha(m, n))$

▶Ref: Tarjan, Data structures and network algorithms, SIAM, 1983

From Darwin to AlphaFold

Intermezzo: Union-Find

AlphaFold predictions at a glance

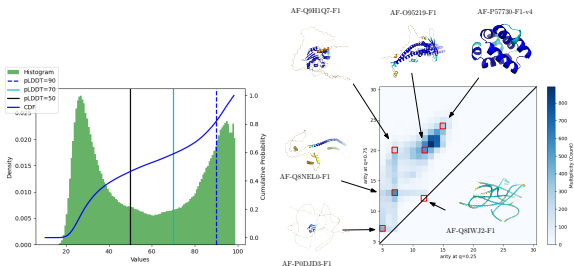
- Packing analysis: methods

- pLDDT: methods

- Results

AlphaFold predictions at a glance:

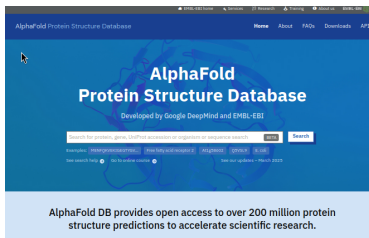
Towards a coherent perspective on
packing properties, pLDDT values, and IDPs/IDRs



F. Cazals, E. Sarti, Centre Inria at Université Côte d'Azur, France

Databases of protein models: AlphaFold-DB

- ▶ **Goal:** provided AlphaFold predictions for all known sequences



<https://alphafold.ebi.ac.uk/>

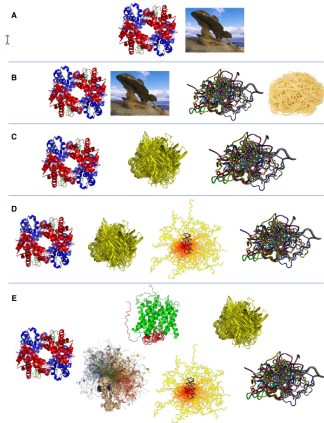
▶ Selected quotes:

- ▶ *"One of the Biggest Problems in Biology Has Finally Been Solved", Scientific American, 2022 <https://www.scientificamerican.com/article/one-of-the-biggest-problems-in-biology-has-finally-been-solved/>*
- ▶ *"... over 200 million protein structure predictions to accelerate scientific research."*

▶Ref: Jumper et al, Nature, 2021

▶Ref: Varadi et al, NAR, 2021

Intrinsically Disordered Regions / Proteins: metaphors

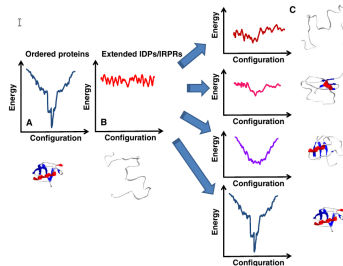
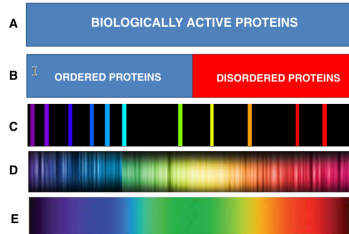


- ▶ (A) singleton: rocks
- ▶ (B) pair: +noodles
- ▶ (C) triplet: + molten globules
- ▶ (D) quartet: + native coil
- ▶ (E) continuum !

▶Ref: Uversky, Unusual biophysics of intrinsically disordered proteins, 2013

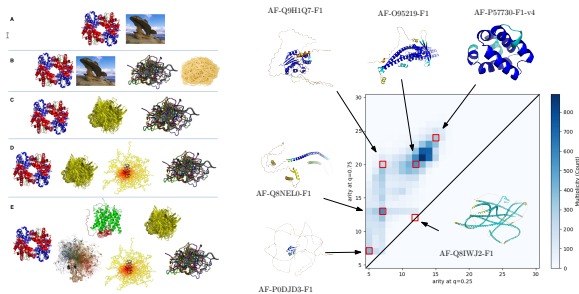
Intrinsically Disordered Regions / Proteins:

energy landscapes and functions



►Ref: Uversky, Unusual biophysics of intrinsically disordered proteins, 2013

AlphaFold-DB: Helping sorting the wheat from the chaff



►Ref: Uversky, Unusual biophysics of intrinsically disordered proteins, 2013

►Ref: Cazals and Sarti,
<https://www.biorxiv.org/content/10.1101/2024.11.16.623929v4>, 2025

Assessing AlphaFold models: main goals

▷ AlphaFold predictions: contacts, domains, whole structures

▶ Methods:

- Unsupervised methods: DCA, MSA transformers
- Supervised: EvoFormer

▶ Our analysis on AlphaFold-DB:

- (Q1) Structures: whole genome analysis and clustering
- (Q2) Domains: high quality vs hallucinations

▷ pLDDT values and order/disorder:

▶ Methods:

- Intrinsically disordered proteins/regions, IDRs and functions

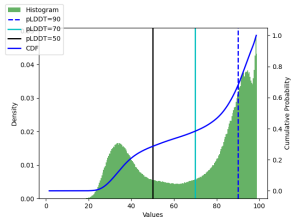
▶ Our analysis on AlphaFold-DB:

- (Q3) pLDDT versus (IDRs / IDPs): true/false positives?
- (Q4) Coherence of pLDDT values along the chain, fragmentation

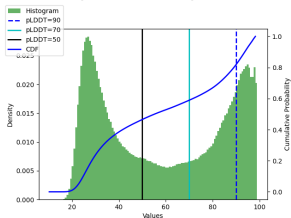
pLDDT values on whole genomes

▷ pLDDT ranges: 100 very high 90 high 70 low 50 very low 0

EXple 1: H. Sapiens



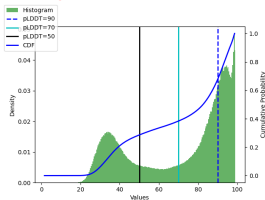
EXple 2: P. Falciparum



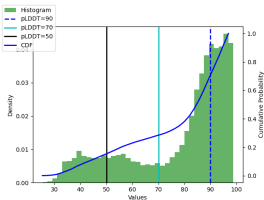
genome / pLDDT	50	70	90
AThaliana/all	0.204	0.315	0.550
CAlbicans/all	0.213	0.312	0.570
CElegans/all	0.202	0.323	0.597
DDiscoideum/all	0.288	0.423	0.681
DMelanogaster/all	0.281	0.387	0.623
DRerio/all	0.246	0.343	0.593
EColi/all	0.029	0.078	0.275
GMax/all	0.217	0.338	0.577
HSapiens/all	0.284	0.382	0.666
MJannaschii/all	0.035	0.084	0.264
MMusculus/all	0.255	0.352	0.597
OryzaSativa/all	0.257	0.408	0.623
RattusNorvegicus/all	0.253	0.351	0.596
SCerevisiae/all	0.213	0.314	0.582
SPombe/all	0.187	0.289	0.558
ZeaMays/all	0.260	0.394	0.635
SAureus/all	0.045	0.096	0.295
HPylori/all	0.053	0.123	0.347
MTuberculosis/all	0.069	0.133	0.322
Aeruginosa/all	0.036	0.086	0.278
PFalciparum/all	0.460	0.584	0.804

pLDDT distributions per genome: illustrations

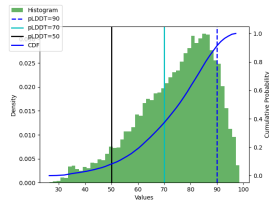
▷ H. Sapiens:



pLDDT all

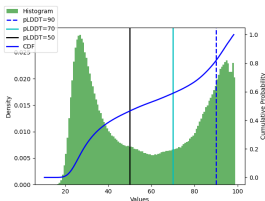


pLDDT median

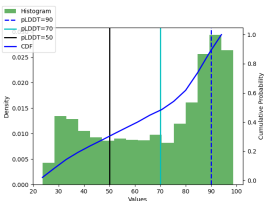


pLDDT mean

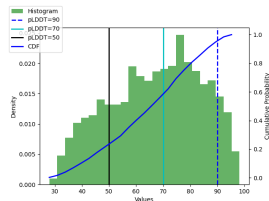
▷ P. Falciparum (cf Malaria):



pLDDT all



pLDDT median



pLDDT mean

Assessing contacts: arity and arity signature

Definition 10. *arity* of a C_α : # neighboring C_α s within distance range $r(= 10\text{\AA})$.

▷ For a polypeptide chain, consider:

- ▶ $L_a = \{a_n, \dots, a_n\}$: arities of the n C_α carbons;
- ▶ $L_A = [A_1, \dots, A_m]$: unique arities sorted by increasing value,
- ▶ CDF_{C_α} : arity CDF – on sorted unique arities in L_A .

Definition 11. (Arity signature $\text{Sig}_{C_\alpha}(P)$ of a polypeptide chain P .)

- ▶ $P_K = \{p_1, \dots, p_K\}$ of increasing percentiles.

Let $\text{arity}_{C_\alpha}(p_k)$: smallest arity A_j such that the CDF is $\geq p_k$:

$$\text{arity}_{C_\alpha}(p_k) = \text{CDF}_{C_\alpha}^{-1}(p_k) \stackrel{\text{Def}}{=} \arg \min_j \text{CDF}_{C_\alpha}(A_j) \geq p_k. \quad (30)$$

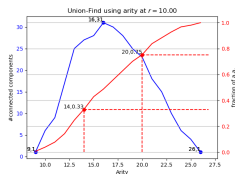
C_α signature: $\text{Sig}_{C_\alpha}(P) = \{\text{arity}_{C_\alpha}(p_1), \dots, \text{arity}_{C_\alpha}(p_K)\}$.

▷ **NB:** with $\{q_1 = 0.25, q_2 = 0.75\}$: the two arity values required to gather 25% and 75% percent of the number of amino acids in the chain.

Arity distribution and arity signature: prototypical folds



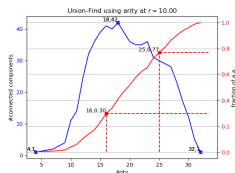
Fold α



Globin, 154 a.a.



Fold β



Propeller, 350 a.a.

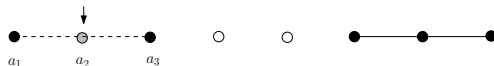
- ▶ Fold α : 101m-globin-alpha; 154 amino acids. Quantile-arity signature [(0.25, 14), (0.5, 17), (0.75, 20)]
- ▶ Fold β : 1erj-propeller-beta; 350 amino acids. Quantile-arity signature [(0.25, 16), (0.5, 20), (0.75, 25)]

Filtrations \mathcal{G}_u of the primary structure

- ▶ **Primary structure:** modeled as a path graph with n nodes/a.a. and $n - 1$ edges
- ▶ **Filtration:** sequence of nested topological subspaces – subgraphs in our case:
 - ▶ *Smallest subspace:* \emptyset
 - ▶ *Largest subspace:* whole path graph *i.e.* primary structure
- ▶ **Construction of a filtration:** each amino acid (vertex) is equipped with a real value u , and a.a. are inserted incrementally

Definition 12. Filtration \mathcal{G}_u : sequence of nested subgraphs obtained by inserting a.a. by increasing value of u .

- ▶ $u = -\infty : \mathcal{G}_u = \emptyset$; $u = \infty : \mathcal{G}_u =$ whole polypeptide chain
- ▶ The num. of c.c. at each value of u is denoted $\nu = N_{cc}(u)$, with $\nu \in [1, \lceil n/2 \rceil]$.



a.a. insertion and reduction of the number of c.c.

▶Ref: R. Tarjan, Data structures and network algorithm, SIAM, 1983

Arity based filtration $\mathcal{G}_{\text{arity}}$

Definition 13. (Arity filtration $\mathcal{G}_{\text{arity}}$) Filtration obtained using as parameter the arity of a C_α .

▷ **Algorithm:** batch processing to handle at C_α with a given arity at once

```
procedure Build_arity_filtration( $\{a_i\}_{i=1,\dots,n}$ )  
  Compute all individual arities  
  Compute the the sorted list  $L = [A_1, \dots, A_m]$  of unique arities  
  for  $A_i \in L$  do  
    for  $c_j \in A_{C_\alpha}^{-1}(A_i)$  do  
      UF.make_set( $c_j$ )  
      if  $j > 1$  and  $c_{j-1}$  exists in the UF data structure: UF.union( $c_j, c_{j-1}$ )  
      if  $j < n$  and  $c_{j+1}$  exists in the UF data structure: UF.union( $c_j, c_{j+1}$ )  
    end for  
     $cc_i \leftarrow$  UF.num_cc  
     $nn_i \leftarrow$  UF.num_nodes  
  end for  
  return  $\{(cc_i, nn_i)\}$  and the associated persistence diagram  
end procedure
```

pLDDT based filtration $\mathcal{G}_{\text{pLDDT}}$

Definition 14. (pLDDT filtration $\mathcal{G}_{\text{arity}}$) Filtration obtained using as parameter $u = -\text{pLDDT}$ values.

NB: increasing the a.a. by increasing $-\text{pLDDT}$ values: high confidence first

▷ **Algorithm:** standard Union-Find

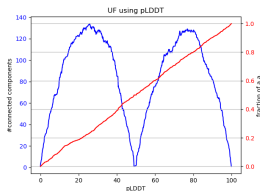
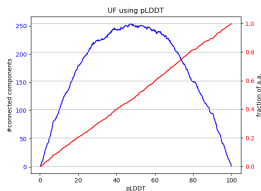
```
procedure Build_path_graph_filtration( $[(j, u_j)]_{j=1, \dots, n}$ )
  Form the list  $[(j, u_j)]$ ,  $j = 1, \dots, n$ , for the  $n$  amino acids
  Let  $L$  be this sorted list ascending  $u_j$  values
  for  $(j, u_j) \in L$  do
    UF.make_set( $c_j$ )
    if  $j > 1$  and  $c_{j-1}$  exists in the UF data structure: UF.union( $c_j, c_{j-1}$ )
    if  $j < n$  and  $c_{j+1}$  exists in the UF data structure: UF.union( $c_j, c_{j+1}$ )
     $cc_j \leftarrow \text{UF.num\_cc}$ 
     $nn_j \leftarrow nn_j + 1$ 
  end for
  return  $\{(cc_i, nn_i)\}$  and the associated persistence diagram
end procedure
```

Union-Find on the polypeptide chain and filtration \mathcal{G}_u . Particular case: using pLDDT as value for the parameter u yields the filtration $\mathcal{G}_{\text{pLDDT}}$.

Filtration $\mathcal{G}_{\text{pLDDT}}$: null model using $u = \text{pLDDT}$

▷ **Rationale:** what is the coherence of pLDDT values along the polypeptide chain?

▷ **Two illustrations:**



- ▶ Left: $n = 1000$ a.a. with random pLDDT values in $[0, 100]$;
- ▶ Right: the first (resp. last) 500 a.a. with random pLDDT values in $[0, 49]$ (resp. $[50, 100]$).
- ▶ (Top row) Blue curve: function $N_{\text{cc}}(\text{pLDDT})$; red curve: fraction of amino acids.
- ▶ (Bottom row) persistence diagrams.

▷ **Conjecture.** For a n -nodes graph/path, the expectation of the maximum of the number of connected components yielded by the incremental construction is equal to $n/4$.

Filtration \mathcal{G}_U : persistence diagram

- *Persistence diagram*: one point per connected component. Persistence of c_i

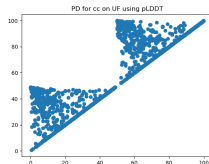
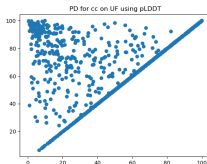
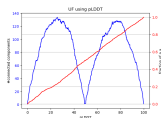
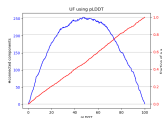
$$p_{\mathcal{G}}(c_i) = \text{death}_{\mathcal{G}}(c_i) - \text{birth}_{\mathcal{G}}(c_i)$$

- *Persistence diagram (PD)*: $P = \{c_1 = (b_1, d_1), \dots, c_m = (b_m, d_m)\}$
- Critical points with persistence > 0 : fraction of positive c.p.:

$$f_{cp}^+ = m'/m.$$

- The *normalized persistence entropy* – assuming a finite set of values \mathcal{P} :

$$H_p = - \sum_{p_i \in \mathcal{P}} \mathbb{P}[p_i] \log \mathbb{P}[p_i] / \log |\mathcal{P}|. \quad (31)$$

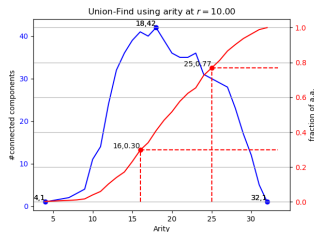


$n = 500 + 500$

$n = 1000$

Filtration \mathcal{G}_u : persistent maximma of the function $\nu = N_{cc}(u)$

► **Goal:** find the salient *salient* / *persistent* local maxima of the function $N_{cc}(u)$.



► **Algorithm sketch to process $N_{cc}(u)$:**

- Filtration on the filtration \mathcal{H}_ν associated with super-level sets $N_{cc}^{-1}([\nu, \lceil n/2 \rceil]$. NB: \mathcal{H} indicates that the filtration is on the height function $N_{cc}(u)$.
- Persistence of a local maximum: elevation drop leading to a saddle point also connected to a more elevated local maximum

$$p_{\mathcal{H}}(c_i) = \text{death}_{\mathcal{H}}(c_i) - \text{birth}_{\mathcal{H}}(c_i)$$

- Simplification of $N_{cc}(u)$: uses the Morse-Smale-Witten chain complex to iteratively cancel pairs of critical points
- Result: $\text{PLM}(t_\nu)$ the number of persistent local maxima of $N_{cc}(u)$ at persistence threshold t_ν .

NB: in practice, with a relative threshold $t_p \in (0, 1)$: $t_\nu = n * t_p$

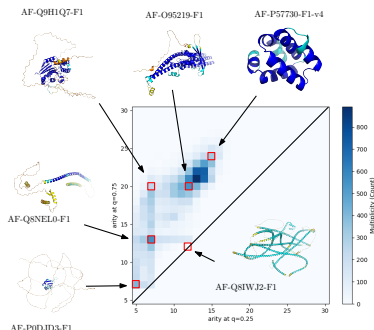
► Ref: Cazals and Cohen-Steiner, Reconstructing 3D compact sets, CGTA, 2011

Arity map of a collection of structures

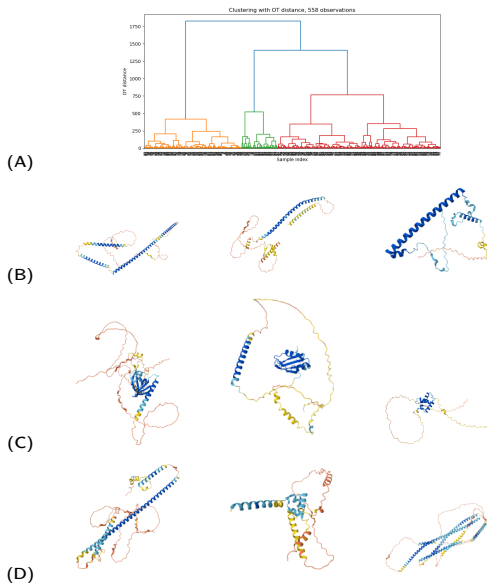
the example of H. Sapiens

▷ **Rationale:** for a collection of structures, perform dimensionality reduction + clustering at once

Definition 15. Given two quantiles $q_1 (= 0.25)$ and $q_2 (= 0.75)$, the *arity map* is the map whose x and y axis are the arities at q_1 and q_2 . A *bin/cell* of the map hosts all structures with a prescribed arity signature.



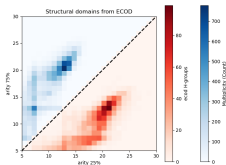
Arity map of HSapiens: clustering in a cell



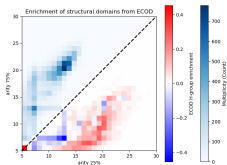
H. Sapiens: hierarchical clustering of the 558 structures in the 7x13 cell of the arity map, with three random structures from every cluster.

Q2. Predicted domains and their quality

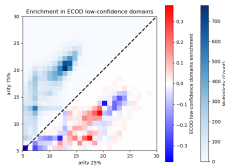
- ▷ **ECOD:** hierarchical classification of protein domains
 - ▷ Top level: close to CATH
 - ▷ Enriched with AlphaFold predictions: high and low quality domains
- ▷ **H. Sapiens:** arity map of the human proteome vs ECOD domain enrichment in human proteins



(A)



(B)

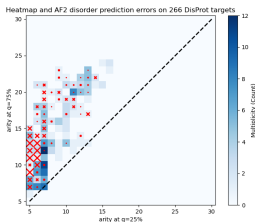


(C)

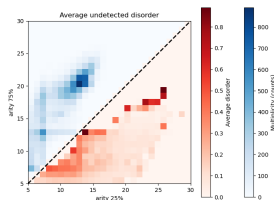
The upper triangle always reports the human proteome arity map for reference. **(A)** Number of unique ECOD H-groups for each arity signature. **(B)** Enrichment difference (with respect to the proteome) of unique ECOD H-groups. **(C)** Enrichment difference (with respect to the proteome) of non-redundant ECOD low-confidence domains.

Q3. Predictions and intrinsically disordered proteins/regions

► H. Sapiens: arity map discriminates IDRs: False positives and False negatives



(A)

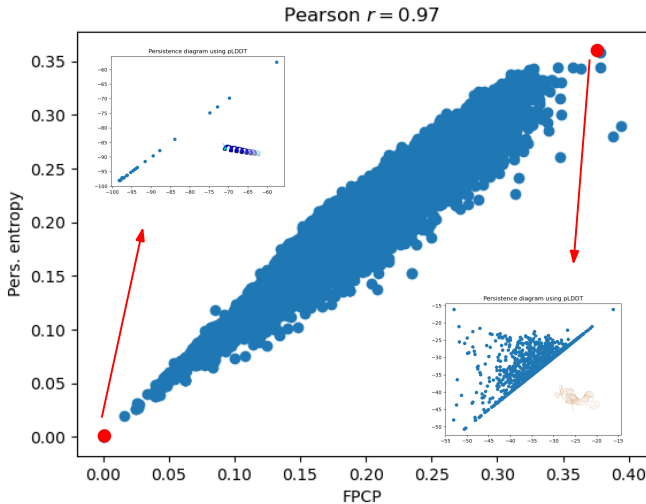


(B)

- **(A) DisProt analysis on 266 target proteins lacking structural data.** The size of the red crosses over the arity map indicates the fraction of amino acids incorrectly predicted as disordered—pLDDT < 50% but lacking a disorder annotation in DisProt.
- **(B) AIUPred analysis.** IDRs on the whole human genome, and for each protein, computation of the number of residues characterized by pLDDT ≥ 0.5 and a disorder score of > 0.5 according to AIUPred. The triangle delimited by $([6, 8], [6, 13], [12, 13])$ is enriched in predicted structures not clearly recognized by AlphaFold.

Q4. pLDDT values and fragmentation of AlphaFold reconstructions

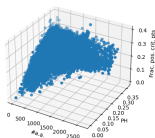
▷ HSapiens predictions and $\mathcal{G}_{\text{pLDDT}}$: regular and random structures



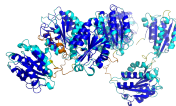
Persistence entropy H_p and fraction of positive critical points (FPCP) f_{CP}^+ : illustrations for HSapiens, with structures achieving the minimum and maximum H_p values.

Q4. pLDDT values and fragmentation of AlphaFold reconstructions

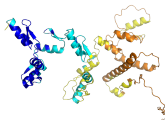
▷ **Fragmentation of HSapiens predictions:** using $\mathcal{G}_{\text{pLDDT}}$, large H_p + large number of local maxima $N_{\text{cc}}(\text{pLDDT})$



(A)



(B; AF-P12111-F4-model-v4)



(C; AF-Q6V9R5-F1-model-v4)



(D; AF-Q9H342-F1-model-v4)

(A) Scatter plot with protein size \times persistence entropy $H_p \times$ fraction of positive critical points f_{cp}^+ . **(B, C, D)** For HSapiens, at persistence threshold $t_p = 0.025$, 86 structures are characterized by $H_p \geq 0.25$, $\#a.a. \geq 200$, $\text{PLM} \geq 3$. Three of them are displayed.

AlphaFold and AlphaFold-DB: take home messages

- ▶ Excellent structures, with high pLDDT all over
- ▶ But a whole zoo, with a number of hallucinations
- ▶ With respect to IDPs/IDRs: both false positives and false negatives

▶Ref: Cazals and Sarti,
<https://www.biorxiv.org/content/10.1101/2024.11.16.623929v4>, 2025



Nicholas Bhattacharya, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K Koo, David Baker, Yun S Song, and Sergey Ovchinnikov.
Single layers of attention suffice to predict protein contacts.
Biorxiv, pages 2020–12, 2020.



Ioan Ieremie, Rob M Ewing, and Mahesan Niranjan.
Protein language models meet reduced amino acid alphabets.
Bioinformatics, 40(2):btac061, 2024.



F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. Marks, C. Sander, R. Zecchina, J. Onuchic, T. Hwa, and M. Weigt.
Direct-coupling analysis of residue coevolution captures native contacts across many protein families.
PNAS, 108(49):E1293–E1301, 2011.



Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt.
Efficient generative modeling of protein sequences using simple autoregressive models.
Nature communications, 12(1):5800, 2021.



Judea Pearl.
Bayesian networks.
2011.



Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker.
Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era.
Proceedings of the National Academy of Sciences, 110(39):15674–15679, 2013.



Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives.
MSA transformer.
In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.



Damiano Sgarbosa, Umberto Lupo, and Anne-Florence Bitbol.
Generative power of a protein language model trained on multiple sequence alignments.
Elife, 12:e79854, 2023.