AlphaFold predictions on whole genomes at a glance

Frédéric Cazals and Edoardo Sarti, Centre Inria at Université Côte d'Azur, France*

December 20, 2024

Abstract

For model organisms, AlphaFold predictions show that 30% to 40% of amino acids have a (very) low pLDDT confidence score. This observation, combined with the method's high complexity, commands to investigate difficult cases, the link with IDPs/IDRs, and potential hallucinations. We do so via four contributions. First, we provide a multiscale characterization of stretches with coherent pLDDT values along the sequence, an important analysis for model quality assessment. Second, we leverage the 3D atomic packing properties of predictions to represent a structure as a distribution. This distribution is then mapped into the so-called *2D arity map*, which simultaneously performs dimensionality reduction and clustering, effectively summarizing all structural elements across all predictions. Third, using the database of domains ECOD, we study potential biases in AlphaFold predictions at the sequence and structural levels, identifying a specific region of the arity map populated with low quality 3D domains. Finally, with a focus on proteins with intrinsically disordered regions (IDRs), we identify another specific region of the arity map enriched for false positives in IDRs.

Summarizing, the arity map sheds light on the accuracy of AlphaFold predictions, both in terms of 3D domains and IDRs.

1 Introduction

AlphaFold: complexity and predictions. AlphaFold is the structure prediction program developed by Deepmind [1]. Enabling the reliable prediction of a protein structure from its amino acid sequence, AlphaFold has opened a new era in structural biology, and its leaders, D. Hassabis and J. Jumper, were co-awarded the 2024 Nobel prize in chemistry. Practically, each prediction is provided as a PDB file, with each amino acid qualified by a reliability score called the pLDDT (prediction local distance difference test) [2]. The AlphaFold developers proposed the following four confidence ranges: very high: $0.9 \leq$ pLDDT; high: $0.7 \leq$ pLDDT < 0.9; low: $0.5 \leq$ pLDDT < 0.7; very low pLDDT < 0.5. On the experimental side, the Protein Data Bank currently features $\sim 230K$ experimental structures against the $\sim 200M$ known protein sequences cataloged in UniProtKB. This disparity underscores the importance of structure prediction for whole whole genomes [3] (SWISS-MODEL Repository), a critical task to advance our understanding of proteomes. This endeavor has recently been undertaken with AlphaFold, and predictions deposited in the AlphaFold-DB database–[4] and https://alphafold.ebi.ac.uk. An exhaustive analysis of structures in AlphaFold-DB shows that for model organisms, AlphaFold predictions contain 30% to 40% of amino acids with a (very) low pLDDT (confidence) score (Section S2).

From the methods' perspective, AlphaFold uses two main modules, known as the Evoformer and the structure module, within which so-called transformers play a key role [5] to unveil long range couplings between amino acids. The overall algorithm is very complex, as evidenced by the ablation study which compares 10 variants to assess the merits of the main building blocks [1, SI]. In the context of the bias-variance trade-off in machine learning in general [6], and of the double descent phenomenon for large deep models [7], an important point is the balance between the number of parameters in AlphaFold and the pieces

*Frederic.Cazals@inria.fr, Edoardo.Sarti@inria.fr

of information used to train it. The number of parameters has been estimated circa 100 millions. On the other hand, the number of non redundant sequences used in the self-distillation procedure is 350k [1, SI, Section 1.3], which, assuming a median size of 500 amino acids, results in about $\sim 1.75 \cdot 10^8$ amino acids. (Distillation is the process of transferring knowledge from a teacher to a student [8], a simplifying process bearing similarities to coreset calculations in a more algorithmic setting learning [9].)

The significant percentage of amino acids with a (very) low pLDDT and the `AlphaFold` complexity call for further investigations to better understand the quality of predictions. We undertake this endeavor using the predictions on whole genomes provided in the database `AlphaFold-DB`.

Contributions. We explore the relationship between `AlphaFold`'s complexity and the relatively high proportion of amino acids with (very) low pLDDT scores by addressing four key questions.

Q0. pLDDT values and fragmentation of `AlphaFold` reconstructions. Quality assessment for `AlphaFold` models is typically done using the pLDDT value provided for each amino acid.

To perform a more accurate model analysis, we provide a multiscale characterization of stretches along the sequence with coherent pLDDT values. This analysis relies on the stability of connected components of amino acids (a.a.) along the protein sequence, upon inserting these (a.a.) by decreasing pLDDT values.

Q1. Predictions and structural motifs. Two related topics in analyzing `AlphaFold` predictions are the pLDDT values and the structural elements produced. The pLDDT metric has been thoroughly analyzed, revealing its effectiveness as an indicator of result robustness; high mean pLDDT values are associated with minimal structural differences between `AlphaFold` and trRosetta models [10]. However, as indicated by a lack of association with B-factors, pLDDT does not appear to correlate with flexibility [11]. Although pLDDT can predict structural disorder, alternative estimators using `AlphaFold`-predicted structures, such as the relative solvent accessibility (RSA), provide a more accurate measure of disorder [12]. In a complementary vein, the structural elements in predictions have been analyzed using the `Geometricus` algorithm [13], which encodes structural fragments based on (i) a decomposition of protein into structural fragments (based on k-mers or spherical neighborhoods), (ii) a discrete encoding of invariants (moments) these fragments into shape-mers, and (iii) a counting step of shape-mers yielding an embedding of the protein. Clustering embeddings (using non-negative matrix factorization) from 21 predicted proteomes from `AlphaFold-DB` yielded 250 clusters, with 20 main clusters/superfamilies [10]. An embedding based on shape-mers has been obtained using dimensionality reduction (t-SNE, [10, Fig. 2]). The analysis proposed by `Geometricus` is very intuitive and is appealing since the encoding can be used to query the database. On the other hand, in the context of `AlphaFold` predictions, each structure is processed as a whole regardless of individual pLDDT values.

To refine structural analysis, we develop a progressive encoding of each structure based on two filtrations, that is a sequence of nested shapes starting with the empty set and ending with the entire structure. Distributions associated to these filtrations can be used to compare structures using optimal transportation theory [14, 15]. Moreover, these distributions can be summarized by a small number of integers, yielding the *arity map*). In 2D, this map yields a natural low dimensional embedding and clustering of all structures.

Q2. Predictions and low quality domains. The aforementioned complexity of `AlphaFold` and the commensurate information size raises the question of potential biases in the model. To address this, we examine the locations of predicted low-quality domains. To understand our approach, it is important to recall that protein domain classification is a cornerstone of protein science [16, 17, 18]. A particularly noteworthy effort is the recently published ECOD structural domain classification of the entire human proteome [19], which leveraged `AlphaFold` predictions to enrich its classification with several new structural domains categorized by confidence scores as high or low. A domain is classified as low-confidence if it appears globular by its predicted aligned error (PAE) description but lacks a strong homologous hit (via a HMM calculation) within the `ECOD` reference–[19] and `prodata.swmed.edu/ecod/`.

We investigate whether the training of `AlphaFold` introduces biases in the prediction of low-quality domains by analyzing their distribution across the arity map. Notably, we identify a specific region of the arity map that is disproportionately populated by such domains.

Q3. Predictions and intrinsically disordered regions. Since its publication, it has been observed that the low-pLDDT regions in `AlphaFold` predictions correlate very well with observed or predicted intrinsic disordered regions (IDRs) [20]. Indeed `AlphaFold` outperforms state-of-the-art methods such as IUPred2 on disordered region detection on a benchmark of `DisProt` targets and PDB controls [10]. Indeed the per-residue pLDDT scores on the `AlphaFold-DB` human proteome prediction yields a bimodal distribution where the two peaks identify ordered and disordered regions [21]. Nonetheless, there are exceptions: in conditionally folded IDRs are usually predicted as stably folded regions, and studies seems to imply that 10% of the low-pLDDT regions are not related to disorder [21].

To complement the previous finding, we show that although most of the `AlphaFold` disorder predictions are supported by experimental evidence, predictions with certain structural characteristics (a specific region in the arity map) are much richer in false positive IDR detection than others.

2 Methods

2.1 Topological analysis

Filtrations and persistence diagrams. We aim to evaluate the *coherence* of the n amino acids of a protein sequence. To do so, we treat the polypeptide chain as a path graph with n vertices and $n-1$ edges. From a topological data analysis perspective [22], we use a filtration $\mathcal{G}_u$ encoding a sequence of nested subgraphs of this path graph. Here, the calligraphic letter $\mathcal{G}$ indicates that the filtration is applied to the path graph. This filtration assumes that each amino acid is equipped with a real value u, and upon inserting a.a. by increasing u value, we connect the a.a. being processed to its neighbors along the sequence if already inserted–Algorithm 1. The filtration starts with the graph containing just one node and ends with the complete path graph. In practice, u could represent the pLDDT score or the number of neighboring C_α carbons within a specified Euclidean distance. As u varies, we track the number of connected components (c.c.) in the graph $\mathcal{G}_u$ using the Union-Find algorithm [23]. This results in a function $\nu = N_{\mathrm{cc}}(u)$ representing the number of connected components at each value of u, with ν in the interval $[1, \lceil n/2 \rceil]$.

The *persistence* of a connected component in $\mathcal{G}_u$ is the time elapsed between its birth and the moment when it merges with a c.c. that was created earlier. The *persistence diagram* (PD) is the diagram whose x-axis (resp. y-axis) if the birth (resp. death) [22]. The points of the PD are denoted $P = \{c_1, \ldots, c_m\}$, and referred to as *critical points*. The persistence of point c_i reads as $\mathtt{p}_{\mathcal{G}}(c_i) = \mathrm{death}_{\mathcal{G}}(c_i) - \mathrm{birth}_{\mathcal{G}}(c_i)$. We also denote m' the number of critical points whose persistence is positive–*i.e.* critical points away from the diagonal, and define the fraction of positive critical points as $f_{cp}^{+} = m'/m$.

Assuming that the persistences belong to a discrete set $\mathcal{P}$, we denote $\mathbb{P}[\cdot]$ the probability mass function $\mathbb{P} : \mathcal{P} \to [0, 1]$. To assess the diversity of persistence values, we consider the normalized entropy of this probability distribution.

Finally, we consider the *salient / persistent* local maxima of the function $N_{\mathrm{cc}}(u)$. We identify such maxima using persistence again on the filtration $\mathcal{H}_\nu$ associated with super-level $N_{\mathrm{cc}}^{-1}([\nu, \lceil n/2 \rceil])$ of the function $N_{\mathrm{cc}}(u)$. (The calligraphic letter $\mathcal{H}$ indicates that the filtration is on the height function $N_{\mathrm{cc}}(u)$.) The persistence $\mathtt{p}_{\mathcal{H}}(c_i) = \mathrm{death}_{\mathcal{H}}(c_i) - \mathrm{birth}_{\mathcal{H}}(c_i)$ of a local maximum of $N_{\mathrm{cc}}(u)$ is the elevation drop leading to a saddle point also connected to a more elevated local maximum. We simplify the function $N_{\mathrm{cc}}(u)$ using the persistence simplification based on the Morse-Smale-Witten chain complex [24]. In particular, we denote $\mathrm{PLM}(t_\nu)$ the number of persistent local maxima of $N_{\mathrm{cc}}(u)$ at persistence threshold t_ν. In practice, we will use a relative threshold $t_{\mathtt{p}}$ to define $t_\nu = n * t_{\mathtt{p}}$ – with n the number of amino acids.

We summarize the previous discussion with the following:

Definition. 1 *Using the filtration $\mathcal{G}_u$ coding subsets of the path graph connecting the amino acids, we define:*

- *The* fraction *of positive critical points:* $f_{cp}^{+} = m'/n$.
- *The* mean persistence *:* $\overline{\mathtt{p}} = \sum_{i=1}^{m} \mathtt{p}_{\mathcal{G}}(c_i)/m$,
- *The* normalized persistence entropy *:* $H_{\mathtt{p}} = -\sum_{\mathtt{p}_i \in \mathcal{P}} \mathbb{P}[\mathtt{p}_i] \log \mathbb{P}[\mathtt{p}_i] / \log|\mathcal{P}|$.

Using the filtration $\mathcal{H}_\nu$ consisting of super-level sets of the function $N_{cc}(u)$, we define:

- *The maximum number of connected components observed:* $N_{cc}^{max} = \max_u N_{cc}(u)$
- *The number of persistent local maxima of the function $N_{cc}(u)$ at persistence threshold t_ν: $PLM(t_\nu)$.*

We now consider two specific filtrations, using the pLDDT quality measure and the number of neighboring C_α carbons in the structure.

2.2 pLDDT based filtration

The filtration $\mathcal{G}_{\text{pLDDT}}$ uses $-$pLDDT as parameter u. We indeed process $-$pLDDT values in increasing order, to favor high quality regions. Assuming that all pLDDT values are different, the PD therefore contains exactly $n-1$ critical points corresponding to the $n-1$ edges connecting the n residues. A number of these have a null persistence and correspond to *accretion*–a C_α carbon inserted in the Union-Find data structure, and merged immediately to a neighbor on the left or the right. These account for the definition of the fraction $f_{cp}^{+} = m'/(n-1)$ (Def. 1).

Example 1 *Consider first a fictitious protein with random pLDDT values in* $[0, 100]$. *The insertion of individual a.a. results in a well behaved curve peaking at the median pLDDT value, and whose maximum is conjectured to be 1/4 of the number of a.a. (Fig. S2; conjecture 1). The persistence diagram also exhibits a characteristic triangular shape with a concentration of critical points near its apex, corresponding to the death of connected components upon moving forward to the full path graph in $\mathcal{G}_{pLDDT}$. Simulations also give the following approximate values:* $n = 100 : f_{cp}^{+} \sim 0.33, H_{\mathtt{p}} \sim 0.36; n = 1000 : f_{cp}^{+} \sim 0.33, H_{\mathtt{p}} \sim 0.45; n = 10,000 : f_{cp}^{+} \sim 0.33, H_{\mathtt{p}} \sim 0.47$.

Consider next the following fictitious protein: (i) a first group of a.a. corresponding to an unstructured region on the N-ter side, with low confidence (say pLDDT $\in [0, 49]$*; and (ii) a second group of a.a. corresponding to a structured region on the C-ter side, with high confidence (say pLDDT* $\in [50, 100]$*). Along the insertion, one observes the formation of the former and the latter regions (Fig. S1).*

2.3 Arity based filtration

Our second filtration is based on the *arity*, which we define first.

Arity signature. A natural parameter to identify compact regions in proteins in the number of neighbors of a C_α carbon. Let the *arity* of a C_α carbon be the number of neighboring C_αs within a distance range r. As classically done, we take $r = 10$Å to account for non-covalent contacts. We note in passing that such neighbors are sufficient to robustly infer protein domains using a direct application of spectral clustering[25].

Assuming the polypeptide chain has n amino acids, define:

- $L_a = \{a_n, \ldots, a_n\}$: arities of the n C_α carbons;
- $L = [A_1, \ldots, A_m]$: unique arities sorted by increasing value;
- $A_{C_\alpha}^{-1}(A_i)$: C_α carbons whose arity is equal to A_i.

The number of C_α with a given arity is used to define the histogram H_{C_α}. We compare two normalized such histograms (unit cumulated sum) using the mass transport solver from [26] implemented in the *Python Optimal Transport* package [27].

To simplify the presentation provided the H_{C_α} and the associated CDF_{C_α}, we define the following signature:

Definition. 2 *(Arity signature $Sig._{C_\alpha}(P)$) Consider the cumulative distribution functions associated with the number of C_α carbons, that is CDF_{C_α}. Consider a list $P_K = \{p_1, \ldots, p_K\}$ of increasing percentiles. For $p_k \in P_K$, let $\texttt{arity_}C_\alpha(p_k)$ be the smallest arity such that the CDF is larger than p_k, that is*

$$\texttt{arity_}C_\alpha(p_k) = CDF^{-1}_{C_\alpha}(p_k) \overset{Def}{=} \arg\min_k CDF_{C_\alpha}(k) \geq p_k. \tag{1}$$

The C_α signature *is the increasing sequence $Sig._{C_\alpha}(P) = \{\texttt{arity_}C_\alpha(p_1), \ldots, \texttt{arity_}C_\alpha(p_K)\}$.*

Using two quantiles $\{q_1 = 0.25, q_2 = 0.75\}$, the arity signature refers to the two arity values required to gather 25% and 75% percent of the number of amino acids in the chain.

For a collection of structures, typically all predictions in a genome, we define:

Definition. 3 *Given two quantiles $q_1(= 0.25)$ and $q_2(= 0.75)$, the* arity map *is the heatmap whose x and y axis are the arities at q_1 and q_2.*

Note that for a given genome, a cell of the arity map hosts all chains with a prescribed arity signature.

Arity filtration $\mathcal{G}_{\texttt{arity}}$. Using the arity as filtration parameter yields the filtration $\mathcal{G}_{\texttt{arity}}$, whose connected components are maintained using Algorithm 2–a simple variant of Algorithm 1 handling all C_α carbons with a given arity in a batch.

3 Results

3.1 Q0. pLDDT values and fragmentation of `AlphaFold` reconstructions

We use the filtration $\mathcal{G}_{\mathrm{pLDDT}}$ and the associated values (Def. 1, Fig. 2) to characterize the coherence of pLDDT values along the backbone.

Specific proteins with different profiles. To appreciate the understanding yielded by our statistics, and in particular the persistence entropy H_{p} and the variation of the number of connected components $N_{\mathrm{cc}}(\mathrm{pLDDT})$, let us consider various protein profiles, namely well structured, disordered, and mixed (Fig. 1). It appears that the function $N_{\mathrm{cc}}(\mathrm{pLDDT})$ varies from unimodal (when considering persistent maxima) to multimodal. The former occurs for well structured or unstructured proteins (Fig. 1(Top, Middle)), and the latter for mixed cases (Fig. 1(Bottom)). For example, the *D. Melanogaster* protein AF-Q9VQS4-F1-model_v4 exhibits a clear bimodal profile with two persistent maxima (Fig. 1(Bottom)). On this example, the plateau at five connected components corresponds to five helices–four long ones plus a shorter one. The persistence diagrams associated with the construction of the filtration also exhibit the dynamics of connected components (Fig. 1, right column). Using the same example, the PD of *D. Melanogaster* quantifies the dynamics of the high (resp. low) confidence region of the molecule on the left (resp. right) part of the diagram. We note that the apex of the triangles are not as populated as those of the null model (Fig. S2), showing a lesser randomness of pLDDT values, which is consistent with the lesser entropy H_{p}.

Let us now analyse these quantities in more detail at the genome scale.

Persistence entropy. At the genome scale, we observe a strong correlation between the three pairs of variables amidst $f_{cp}^{+}, H_{\mathtt{p}}, \overline{\mathtt{p}}$ (Fig. 2(First column), Fig. S3). The maxima of f_{cp}^{+} are consistent with those a the null random model, while those of $H_{\mathtt{p}}$ are slightly below those of the same null model. (NB: average values of $\overline{\mathtt{p}}$ are irrelevant for the null model.) For all organisms, the minimum $H_{\mathtt{p}}$ is observed for very regular structures for which all pLDDT values are within a small range (Fig. 2(Second column)). Instead, the maxima are observed for highly disordered structures whose a.a. span a wide range of pLDDT values (Fig. 2(Third column)). For such cases, the typical triangular shape in the PD is also observed. In considering the overall pLDDT statistics per organism (Table S1), it appears that shifting the pLDDT values towards 100 result in a lesser fraction of positive critical points and in a smaller entropy (Fig. S3).

Maximum value $N_{\mathbf{cc}}^{max}$ of the function $N_{\mathbf{cc}}$(pLDDT). The maximum number of local maxima observed N_{cc}^{max} normalized by the protein size (Fig. S4) in general lower than that of the null model – conjectured to be equal to 1/4, even though some cases reach or even surpass that bound. This is accounted for by two reasons. First, accretion is at play–see definition above. Second, pLDDT values come in batches, so that local insertions into *stretches* along the sequence prevent the maximum number of local maxima to rise. These batches are identified by the persistent local maxima discussed next.

Persistent local maxima PLM of the function $N_{\mathbf{cc}}$(pLDDT). We first note that the number of persistent local maxima is in general less than 10 (Fig. S3(Third column)). Also, the correlation between $H_{\mathtt{p}}$ (or f_{cp}^{+}) and PLM is quite low, which is expected. One can indeed have a significant number of persistent critical points localized in a narrow range of pLDDT values (Fig. S5). This phenomenon is well known to geographers defining peaks on mountains: two tall peaks can be separated by a low lying pass, but be located at a short flying distance [28]. Nevertheless, as pointed out above, persistent local maxima identify coherent stretches along the sequence in terms of pLDDT values.

Model assessment and database queries. In the model assessment perspective, querying the encoding provided by the number of a.a., the persistence entropy $H_{\mathtt{p}}$, the fraction of positive critical points f_{cp}^{+} and the number of persistent local maxima PLM enables the identification of cases incurring a large fragmentation. Such cases exhibit a wide range of pLDDT values aggregated in coherent stretches along the sequence, yielding a large entropy and a (relatively) high number of persistent local maxima of the function N_{cc}(pLDDT) (Fig. 3).

Remark 1 *For additional examples, the reader is referred to the following figures, panel (E): folded proteins: Fig. S8, Fig. S9; unfolded proteins: Fig. S10, Fig. S11; proteins with folded and unfolded regions: Fig. S12, Fig. S13.*

3.2 Q1. Predictions and structural motifs

Charting whole genomes with the arity map. We provide a simple and readable encoding of these structural elements using the *arity* of a C_α carbon, namely the number of neighbors within a ball of fixed radius – taken as 10Å to account for non covalent contacts. In regions with classical α and β folds, we expect the arity mode to be in the range 15-20 (Section S1, Fig. S7). In unstructured regions, the mode is typically in the interval 5-7. For a given prediction, having sorted all C_αs by increasing arity value, we compute the corresponding histogram H_{C_α} and associated cumulative distribution function denoted CDF_{C_α}. The histogram provides a detailed encoding of packing properties of all residues, and two histogram can be compared using optimal transportation theory in general, and the so-called *Earth mover distance* in particular [14, 15]. This histogram can be summarized using a few specific values. Consider two quantiles, say $q_1 = 0.25$ and $q_2 = 0.75$. We define the *arity signature* of a polypeptide chain as the minimum arities for the CDF to reach q_1 and q_2 respectively, which we denote as $\mathtt{arity_}C_\alpha(q_1) = \mathrm{CDF}_{C_\alpha}^{-1}(q_1)$ and $\mathtt{arity_}C_\alpha(q_2) = \mathrm{CDF}_{C_\alpha}^{-1}(q_2)$ (Fig. 4). We process all proteins of a given genome and build the *arity map*, namely the heatmap counting the number of predictions with a prescribed arity signature (H. Sapiens, Fig. 5).

The analysis of this map provides a whole genome analysis at a glance. The x-axis `arity_`$C_\alpha(q_1)$ characterizes the extent of *loose packing*–low values $\mathrm{CDF}^{-1}_{C_\alpha}(q_1)$ is an indication of the presence of unstructured regions, while the y-axis characterizes the degree of *compactness*. For a given cell of this map, the distance to the diagonal $y = x$ encodes the coherence of the overall packing of the region of the protein selected by the quantiles q_1 and q_2. For example, for square on the diagonal, a fraction $q_2 - q_1$ of a.a. have the same arity, stressing the coherence of the overall 3D structure.

We produced the arity map for all models genomes from `AlphaFold-DB` (H. Sapiens, Fig. 5; other species, Section S4). At first sight, these maps are bimodal. The most pronounced mode corresponds to well folded proteins, in the arity map region $[14, 18] \times [20, 25]$. The second mode corresponds to the region $[4, 8] \times [10, 15]$, with in particular a dark spot in the cell 7×13. These modes are in general separated by low density regions in the heat map. The inspection of genomes beyond the model ones yields different patterns, though. For example, P. Falciparum exhibits a continuum between the previous two modes, which is consistent with the overall low quality of predictions noted above.

Notably, arity and pLDDT are found to strongly correlate (Fig. 5(B)), confirming that the arity is also a good proxy for prediction robustness. Yet, pLDDT is not able to show the finer details of the structural repartition that emerges from the arity map.

We further investigate the hot spot at 7×13 of the human genome (Fig. 6). Contrary to the rest of the $[4, 8] \times [10, 15]$ region, this hot spot does not find any counterpart neither in the arity plot of the human proteins of known structure (Fig. 5(C), in red) nor in the arity plot of experimentally-confirmed disordered proteins (Fig. 8(A)). These `AlphaFold-DB` predictions therefore represent disordered structures or inaccurate predictions. We also ran the state-of-the-art software `AIUPred` for predicting IDRs on the whole human genome, and for each protein we computed the number of residues that were associated with both a sufficient confidence score according to `AlphaFold`'s pLDDT and a disorder score of > 0.5 according to `AIUPred` (Fig. 8(A)). The triangle delimited by $([6, 8], [6, 13], [12, 13])$ is enriched in predicted structures where predicted IDRs were not clearly recognized by `AlphaFold`. We note that this region is somehow complementary to the one highlighted by `DisProt`, and that the only neighborhood that both displays false positive and false negative IDR prediction is centered in 7×13.

To further characterize these proteins, we perform a hierarchical clustering of arity histograms using the Earth Mover distance as metric [14, 26]. As already noticed, a 2D arity map reduces the arity histograms using two integers only. As expected, the clustering unveils the finer structure of proteins in this cell, with three main classes (Fig. 6(B,C,D)).

Arity based filtration. We now sketch the topological analysis methods used to bridge the gap between pLDDT values and structural aspects. See Section S1 for the details.

The variability of arities coding the compactness of the atomic environments around C_αs allows incrementally *building* the protein by inserting the a.a. by order of increasing arity and connecting each a.a. to its predecessor (N-ter) and successor (C-ter) if already inserted. This process, which again consists of incrementally inserting a.a., yields a sequence of nested shapes called a *filtration* [22]. Along the way, one maintains the connected components which appear and merge, using a Union-Find data structure [23]. The curve coding the evolution of the number of c.c. is termed the *connected components with arity* curve (c.c. curve for short). We also assume one is able to detect the *salient* (also called *persistent*) local extrema on these curves.

The case of well folded proteins follows this pattern as focusing on salient extrema yields a unimodal arity curve with a salient extremum in the range 15-20 (Fig. 4(Top, blue curve)). The same holds for unstructured proteins, excepts that the local extremum is attained at values $\sim 5 - 7$ (Fig. 4(Middle, blue curve)). The signature is more informative for proteins having both high and low quality regions (Fig. 4(Bottom, blue curve)). For such proteins, the curve of connected components is typically bimodal, since one successively forms the regions with low and high arity.

Concluding, the ability to detect the persistent local extrema on these curves gives a direct indication on the structural elements encountered in the predictions.

Remark 2 *For additional examples, the reader is referred to the following figures, panel (C): folded proteins: Fig. S8, Fig. S9; unfolded proteins: Fig. S10, Fig. S11); proteins with folded and unfolded regions: Fig. S12, Fig. S13).*

3.3 Q2. Predictions and low quality domains

We investigate potential learning biases in `AlphaFold` predictions at the sequence and structure levels.

At the sequence level, counting the number of homologous structures in `UniProtKB` for each arity signature exhibits a clear trend between the arity signature and the average number of homologous sequences (Section S1.3, Fig. S38(A)). However, mapping these homologs to their cluster representatives in the current release of `Uniclust30` erases this bias (Fig. S38(B)), simply showing that structured proteins have more homologous sequences in nature. We also note that the quality of predictions in terms of average pLDDT shows a gradient towards high arity values (Fig. 5(B)). Since both the `AlphaFold` direct and distillation training algorithms construct MSAs of non-redundant sequences with sequence identity $< 30\%$, we conclude that a large and diverse input MSA is necessary but not sufficient for guaranteeing high-quality predictions.

We proceed with the investigation of potential biases at the structural level using the recently published `ECOD` structural domain classification of the entire human proteome [19]. Using `AlphaFold-DB` predictions, `ECOD` could enrich its classification with several new structural domains. At the level of H-groups in `ECOD` hierarchy, each identifier refers to a cluster of sequence-homologous, same-topology domains in tight relation with the SCOP superfamily level. For each arity signature, we thus collect the unique `ECOD` H-group identifiers and the corresponding folds. The resulting 2D-histogram shows a clear correlation with the protein count and the number of `ECOD` folds (Fig. 7(A), upper triangle for count and lower triangle for `ECOD` folds). In order to remove the correlation and study the residual signal, we consider the signed difference of the two rescaled histograms. (The rescaling ensures that the two maxima are equal.) This difference map shows two main domains (Fig. 7(B), red and blue regions). On the one hand, the region of the main peak *i.e.* 13×20 until the 7×13 peak (excluded) shows an enrichment in `ECOD` domains. On the other hand, the region of the 7×13 peak towards the bottom left of the arity map (IDP region, see next Section) gets depleted.

`ECOD` also provides a set of low-confidence domains associated with low pLDDT and PAE (predicted aligned error) values [19], enabling the calculation of the enrichment relative to this dataset. Remarkably, the former *red* region gets split, the enrichment in low confidence domains being confined to the 7×13 signature as well as the arity signatures $[5, 15] \times [14, 20]$. This extended region contains structures richer in IDRs, but still having well-formed domains. Conversely, the $[5, 7] \times [5, 12]$ region is poor both in standard and low-confidence well-formed domains, hinting at a complete dominance of IDRs over structured regions. These observations call for a further inspection of IDPs/IDRs in the arity map.

3.4 Q3. Predictions and intrinsically disordered regions (IDRs)

Several recent works assume that in `AlphaFold`'s predictions, regions with low pLDDT and/or not displaying secondary nor tertiary structure are to be considered intrinsically disordered rather than just the product of insufficient or unexploitable information. The tools developed in the previous sections can help assessing whether and when a correlation between unstructured regions in `AlphaFold`'s prediction correlate with experimental evidences of disordered regions. To do so, we consider the `DisProt` database, which collects disordered regions from more than 1300 target protein sequences. Many of the benchmarked sequences correspond to proteins of known structures, present in the PDB. For this test, we focus on the 266 targets lacking structural data and that have a predicted structure on `AlphaFold-DB`. We take a threshold of pLDDT $< 50\%$ to define low confidence regions and check whether they correspond to annotated IDRs in `DisProt`. Notably, `DisProt` is known not to provide complete annotations, thus in principle `AlphaFold` could discover new unannotated IDRs on the `DisProt` targets. The simplest hypothesis consists of expecting an incompleteness proportional to the extent of the predicted disordered regions–regardless of other structural characteristics of `AlphaFold` predictions. However, while `AlphaFold` predicts a large number of amino acids as part of IDRs for the region of the arity map $\texttt{arity_}C_\alpha(0.25) < 7$, most of these are not annotated as such

in DisProt (Fig. 8). Predictions populating that part of the arity plot are thus likely to be of poor quality: this is not surprising, as the previous section details that proteins in this region often have low numbers of homologous structures.

4 Discussion and outlook

This work introduces a novel representation of structural elements in proteins by employing two filtration of amino acids.

The first one uses pLDDT values along the sequence. The filtration and the associated statistics enable the characterization of cases with a wide range of pLDDT values aggregated in coherent stretches along the sequence, yielding a large entropy and a (relatively) high number of persistent local maxima of the function counting connected components. These analysis make it possible, in particular, to isolate high quality regions of a model.

The second filtration depends on the 3D packing properties of atomic environments, and the corresponding cumulative distribution counting the number of C_α carbons with a prescribed number of neighbors. This distribution enables the comparison of structures using optimal transportation, and the persistent components of the filtration can be used to compare structural motifs from different proteins. To simplify this representation, the distribution can be reduced to two integers, forming the *arity map*, which simultaneously performs clustering of coherent structures and dimensionality reduction. At the genome-wide level, this clustering is used to study two important properties. First, we identify a specific region of the arity map populated with low quality structures. The region is enriched in low-confidence domains recently published in the AlphaFold enriched version of the ECOD database. Most of these domains may be termed *hallucinations* since they may appear valid, but are incorrect and/or fabricated. Second, we use the arity map to study predicted intrinsically disordered regions, showing that it exhibits another specific region with a high rate of false positive IDRs. While the performances of AlphaFold for IDP prediction have been praised, the notion of isolated structure for IDPs is iconoclastic since dynamics play a key role, and future work will likely address the generation of ensembles. We anticipate that these analysis will prove useful to understand limitations of AlphaFold regarding low quality domains and false positive IDRs.

Overall, we believe that the statistics developed in this work complement those already delivered within AlphaFold-DB.

Acknowledgments. This work has been supported by the French government, through the 3IA Côte d'Azur Investments (ANR-19-P3IA-0002), and the ANR project Innuendo (ANR-23-CE45-0019).

We wish to thank V. Mallet for insightful comments.

References

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[2] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[3] Stefan Bienert, Andrew Waterhouse, Tjaart AP De Beer, Gerardo Tauriello, Gabriel Studer, Lorenza Bordoli, and Torsten Schwede. The swiss-model repository—new features and functionality. *Nucleic acids research*, 45(D1):D313–D319, 2017.

[4] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure

database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

[5] Bernard Moussad, Rahmatullah Roche, and Debswapna Bhattacharya. The transformative power of transformers in protein structure prediction. *Proceedings of the National Academy of Sciences*, 120(32):e2303499120, 2023.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of statistical learning : data mining, inference and prediction*. Springer, 2001. htf-esldm-01.

[7] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[8] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[9] A. Maalouf, G. Eini, B. Mussay, D. Feldman, and M. Osadchy. A unified approach to coreset learning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(5), 2024.

[10] Mehmet Akdel, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H. M. Rodrigues, Alistair S. Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Jérôme Basquin, Kresten Lindorff-Larsen, Alex Bateman, Andrey V. Kajava, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B. Ascher, Janet M. Thornton, Norman E. Davey, Amelie Stein, Arne Elofsson, Tristan I. Croll, and Pedro Beltrao. A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, November 2022.

[11] Oliviero Carugo. PLDDT values in AlphaFold2 protein models are unrelated to globular protein local flexibility. *Crystals (Basel)*, 13(11):1560, November 2023.

[12] Damiano Piovesan, Alexander Miguel Monzon, and Silvio C E Tosatto. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci.*, 31(11):e4466, November 2022.

[13] Janani Durairaj, Mehmet Akdel, Dick de Ridder, and Aalt DJ van Dijk. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, 36(Supplement_2):i718–i725, 2020.

[14] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[15] C. Villani. *Topics in optimal transportation*. Number 58. AMS, 2003.

[16] Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E Brenner, Tim JP Hubbard, Cyrus Chothia, and Alexey G Murzin. Data growth and its impact on the scop database: new developments. *Nucleic acids research*, 36(suppl_1):D419–D425, 2007.

[17] Ian Sillitoe, Natalie Dawson, Tony E Lewis, Sayoni Das, Jonathan G Lees, Paul Ashford, Adeyelu Tolulope, Harry M Scholes, Ilya Senatorov, Andra Bujan, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic acids research*, 47(D1):D280–D284, 2019.

[18] Hua Cheng, R Dustin Schaeffer, Yuxing Liao, Lisa N Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V Grishin. ECOD: an evolutionary classification of protein domains. *PLoS computational biology*, 10(12):e1003926, 2014.

[19] R Dustin Schaeffer, Jing Zhang, Kirill E Medvedev, Lisa N Kinch, Qian Cong, and Nick V Grishin. ECOD domain classification of 48 whole proteomes from AlphaFold structure database using DPAM2. *PLoS Comput. Biol.*, 20(2):e1011586, February 2024.

[20] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, August 2021.

[21] T. Reid Alderson, Iva Pritišanac, Đesika Kolarić, Alan M. Moses, and Julie D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by alphafold2. *Proceedings of the National Academy of Sciences*, 120(44), October 2023.

[22] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. AMS, 2010.

[23] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

[24] F. Cazals and D. Cohen-Steiner. Reconstructing 3D compact sets. *Computational Geometry Theory and Applications*, 45(1-2):1–13, 2011.

[25] F. Cazals, J. Herrmann, and E. Sarti. Simpler protein domain identification using spectral clustering. *Under revision*, NA(NA), 2024.

[26] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[27] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[28] C. Thöni. Criteria to define summits in the Swiss alps: prominence and culminance height. *Les Alpes*, 1:26–28, 2003.

5 Artwork

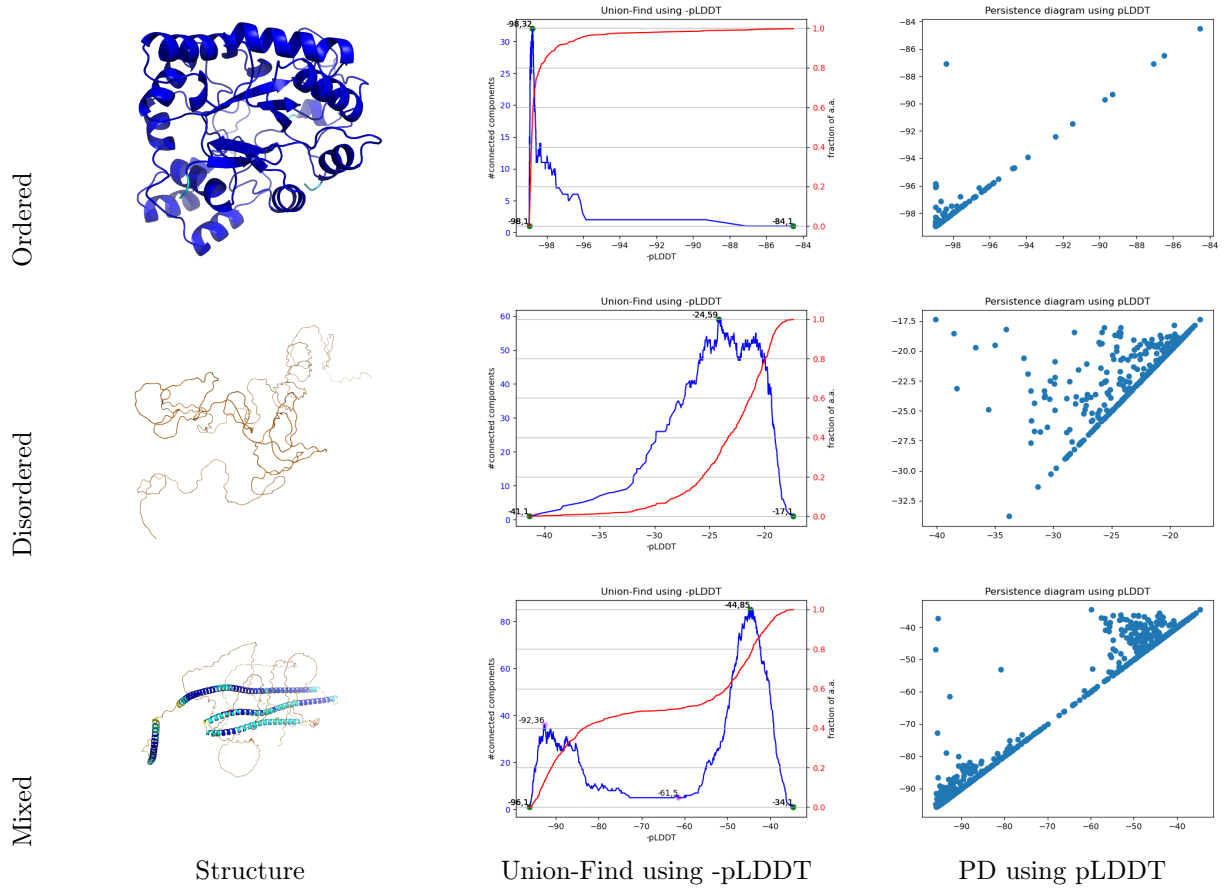


Figure 1: **pLDDT based filtration and evolution of the number of connected components on three prototypical examples. (Top)** (H. Sapiens; AF-P15121-F1-model_v4; 316 amino acids; $H_p \sim 0.04$.) **(Middle)** (Zebrafish; AF-A0A0G2L439-F1-model_v4; 449 amino acids; $H_p \sim 0.27$.) **(Bottom)** (D. Melanogaster; AF-Q9VQS4-F1-model_v4; 781 amino acids; $H_p \sim 0.28$.) Left column: the blue curve codes the variation of the number of connected components; the red one is the CDF of the number of a.a. processed at a given x-value. Right column: persistence diagram for connected components. See also Fig. 4.

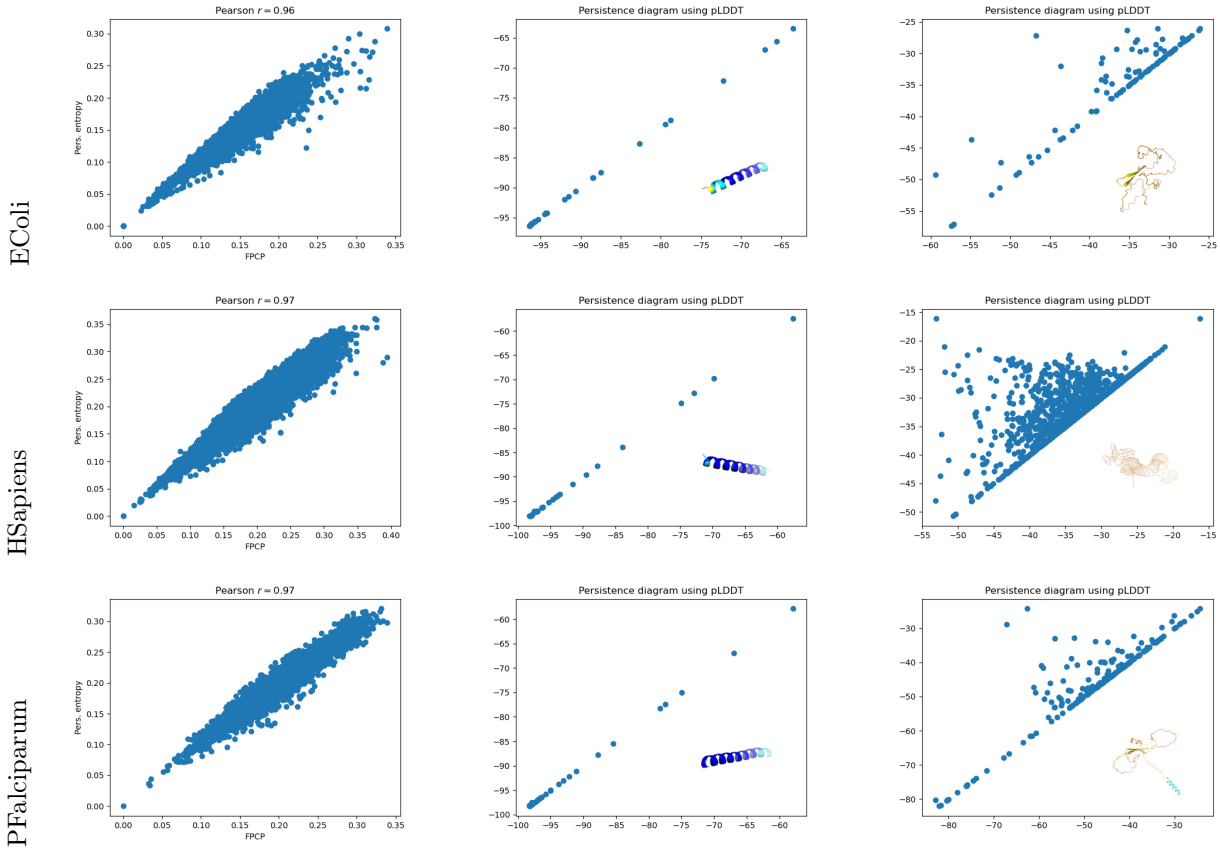
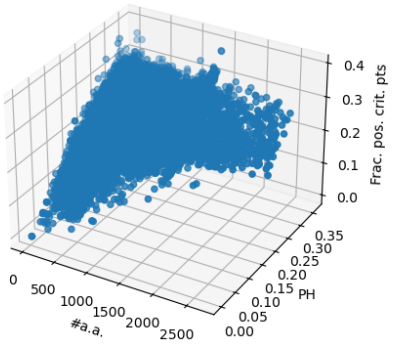
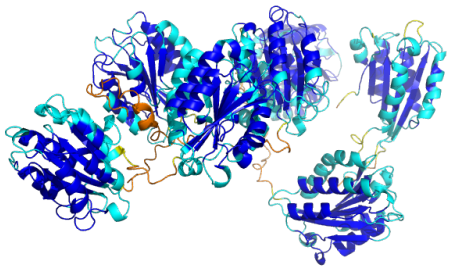


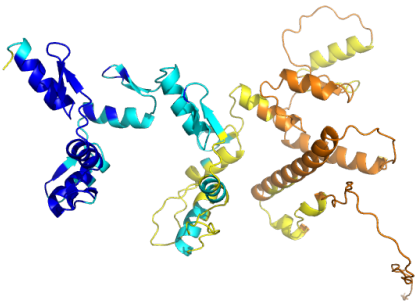
Figure 2: **Persistence entropy H_{p}: illustrations per genome, with structures achieving the minimum and maximum values.** Compare with Table S1. **(First column)** Scatter plot fraction of positive critical points f_{cp}^{+} x persistence entropy H_{p}. **(Second columns)** Per organism, `AlphaFold` model achieving the minimum entropy H_{p}. **(Third columns)** Per organism, `AlphaFold` model achieving the maximum entropy H_{p}. **(Top row: EColi)** AF-A5A617-F1-model_v4, 27 a.a.; AF-P62066-F1-model_v4, 106 a.a. **(Middle row: HSapiens)** AF-O00631-F1-model_v4, 31 a.a.; AF-Q02817-F15-model_v4, 1400 a.a.; **(Bottom row: PFalciparum)** AF-A0A5K1K906-F1-model_v4, 32 a.a.; AF-Q8IM51-F1-model_v4, 202 a.a.



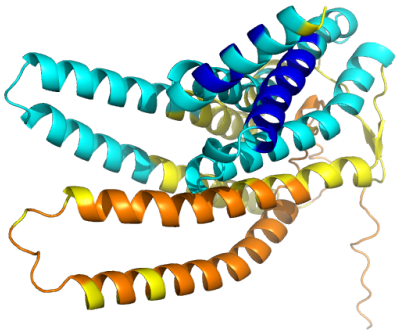
(A)



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



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



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

Figure 3: **The HMA diagram identifies models with high fragmentation: wide range of pLDDT values aggregated in coherent stretches along the sequence.** (A) Scatter plot with protein size x persistence entropy H_{p} x fraction of positive critical points f_{cp}^{+}. For HSapiens, at persistence threshold $t_{\mathrm{p}} = 0.025$, 86 structures are characterized by $H_{\mathrm{p}} \geq 0.25, \#a.a. \geq 200, \mathrm{PLM} \geq 3$. Three of them are displayed.

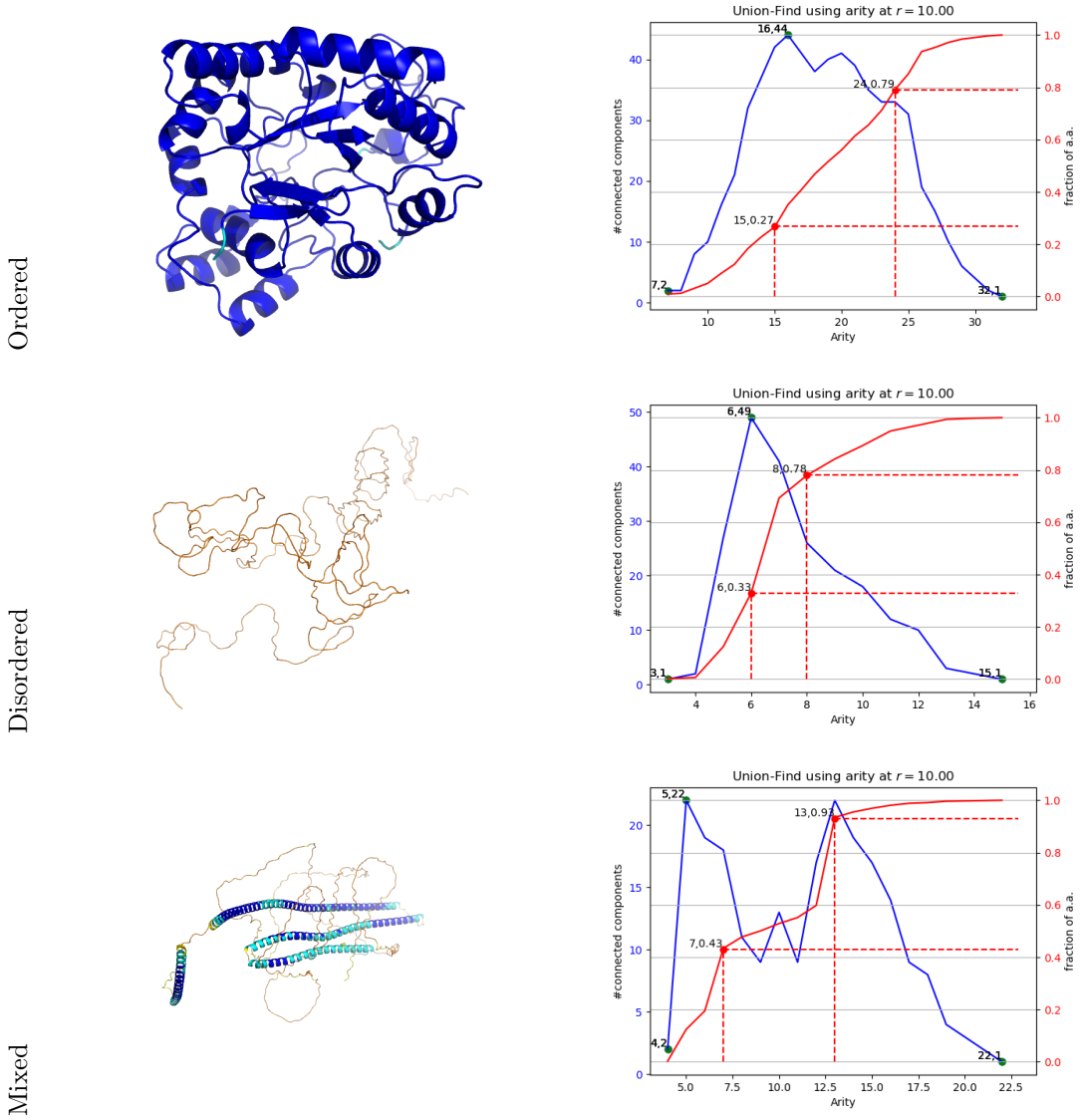


Figure 4: **Arity signature and arity based filtration on three prototypical examples.** Each structure is represented in cartoon mode; the blue (resp. red) a.a. is that featuring the lowest (resp. highest) pLDDT value. The **arity filtration** is the evolving graph defined by the connected components of the polypeptide chain induced by all C_α carbons whose arity is less than the evolving threshold. The associated **connected components curve** (blue curve, left scale) encodes the variation of the number of connected components in the filtration, with salient extrema as blue bullets. The **arity cumulated distribution function** (red curve, right scale) represents the cumulated fraction of C_α carbons whose arity is less than a threshold. The **arity signature** is defined by the interval of arity values required to gather fractions $q_1(= 0.25)$ and $q_2(= 0.75)$ of amino acids – red bullets on the CDF curve. **(Top)** Protein with well folded regions (H. Sapiens; AF-P15121-F1-model_v4; 316 amino acids.) Quantile-arity signature [(15, 0.25), (19, 0.5), (24, 0.75)]. **(Middle)** Protein essentially disordered. (Zebrafish; AF-A0A0G2L439-F1-model_v4; 449 amino acids.) Quantile-arity signature [(6, 0.25), (7, 0.5), (8, 0.75)] **(Bottom)** Protein with ordered and disordered regions. (D. Melanogaster; AF-Q9VQS4-F1-model_v4; 781 amino acids.) Quantile-arity signature [(0.25, 7), (0.5, 9), (0.75, 13)].

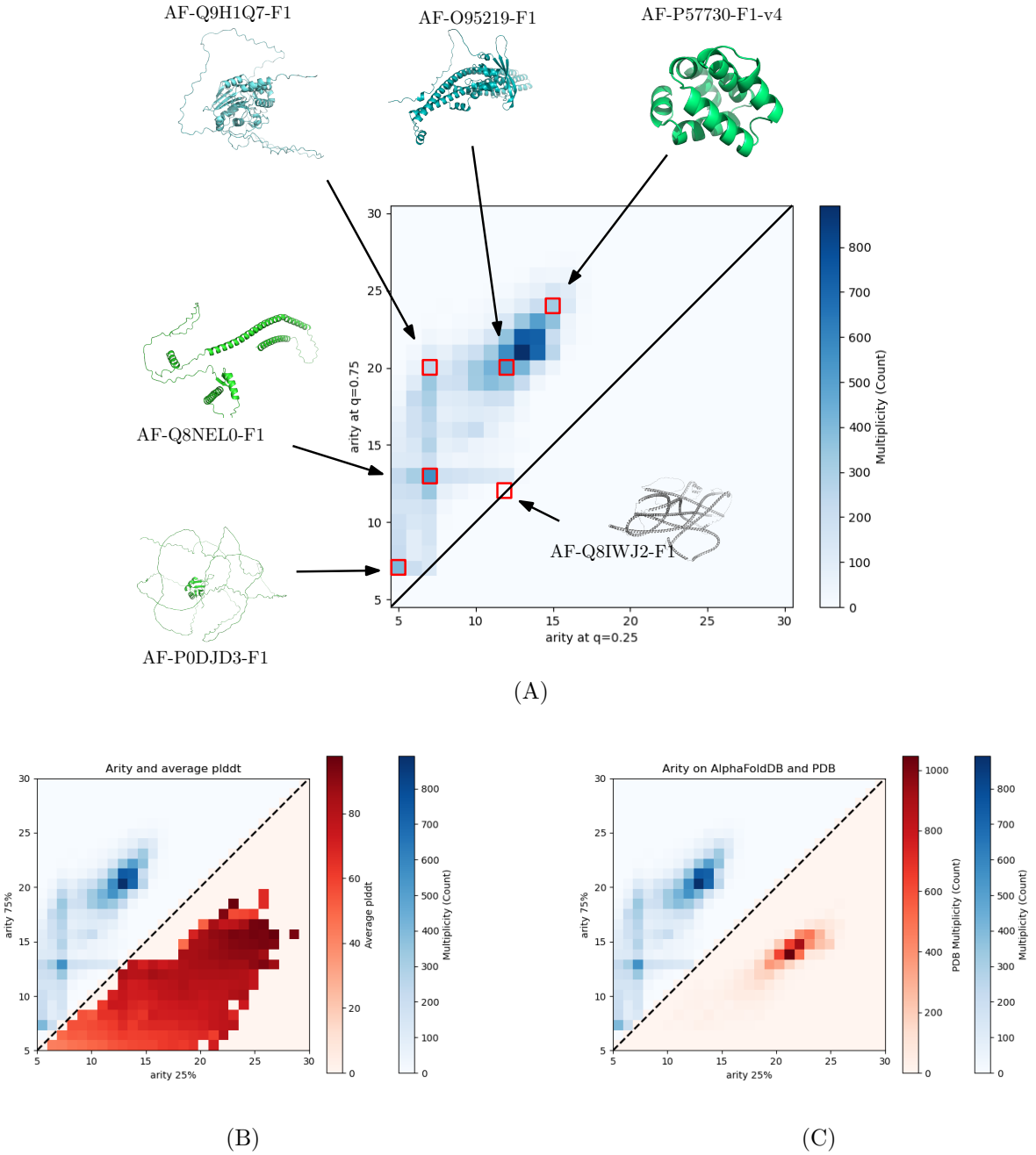


Figure 5: **H. Sapiens: arity map. (A)** Each bin corresponds to all `AlphaFold` predictions harboring a given arity signature. The upper triangle of the heatmap shows the protein count with for a prescribed arity signature. **(B)** Lower triangle: average pLDDT per bin, with mirrored coordinates. Arity and pLDDT are strongly correlated: the highes the first and last quartile are, the highest the averge pLDDT is. Arity is thus a good estimator of structural complexity. **(C)** Lower triangle: protein count from the PDB. Well-formed, ordered proteins only populate the (15,20) peak of the arity plot, confirming that the other two peaks visible in the arity plot of the `AlphaFold-DB` predictions must represent disordered structures or inaccurate predictions.

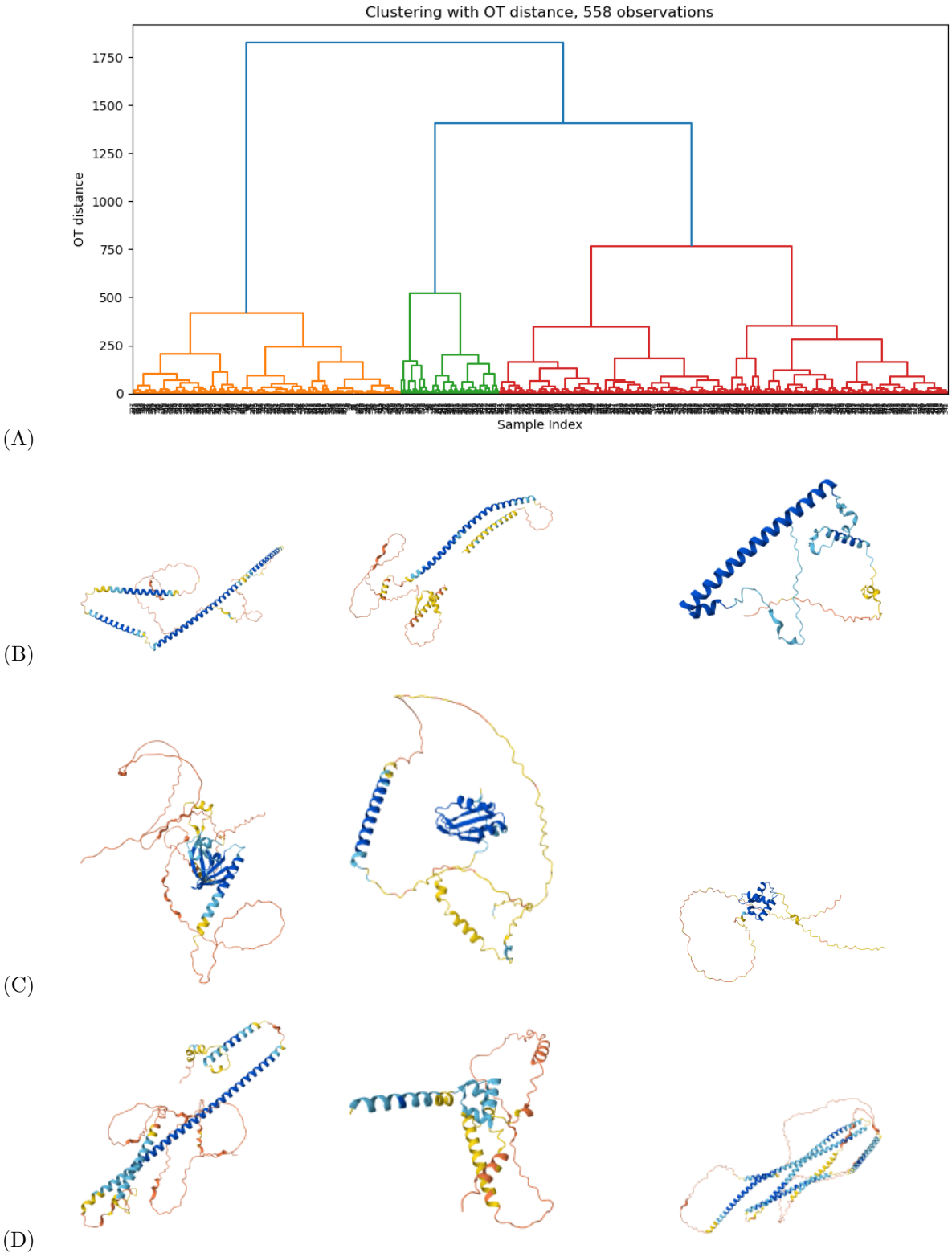
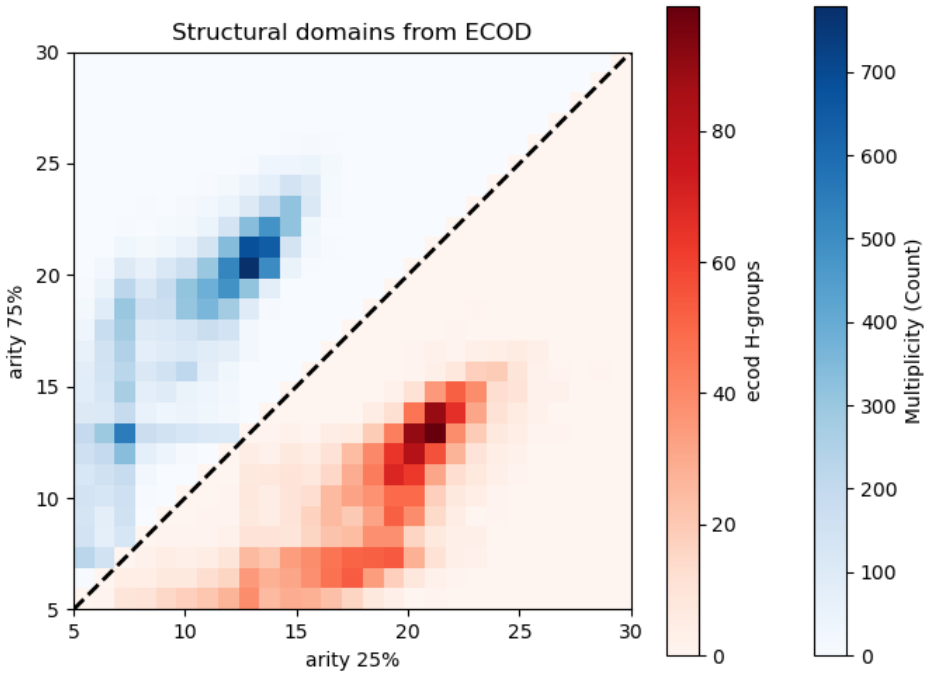
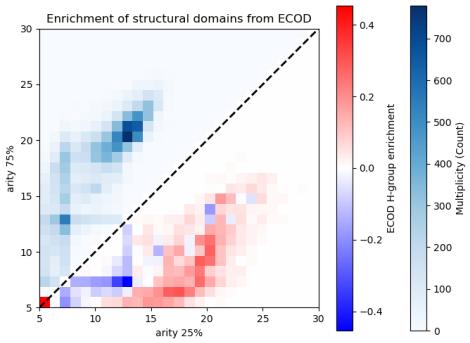


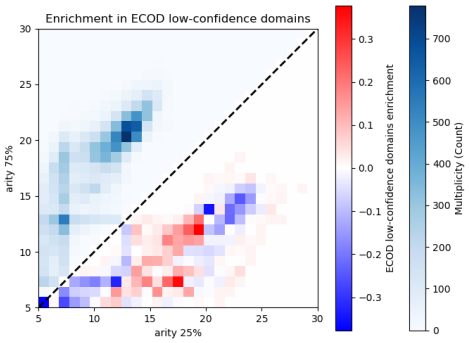
Figure 6: **H. Sapiens: hierarchical clustering of the 558 structures in the 7x13 cell of the arity map, with three random structures from every cluster. (A)** Ward hierarchical clustering of the structures using the optimal transportation distance on the normalized arity histograms H_{C_α} for each pair. **(B)** First cluster independent helices; **(C)** Second cluster: independent helices and a small β core; **(D)** Third cluster: packed helices.



(A)

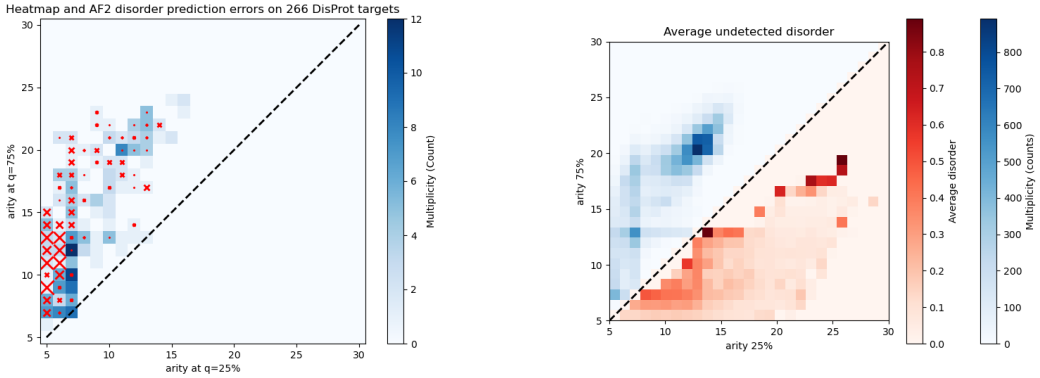


(B)



(C)

Figure 7: **H. Sapiens: arity map of the human proteome vs ECOD domain enrichment in human proteins**. 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.



(A) (B)

Figure 8: **H. Sapiens: disagreement between pLDDT and disorder annotations and predictions**. **(A)** The size of the red crosses over the heatmap indicates how frequently a pLDDT of less than 50 does not identify an amino acid in an experimentally annotated disordered region (no cross: never, pizel-size cross: more than 70% of the times. The leftmost region of the plot shows a high enrichment in such cases. **(B)** The upper triangle reports the human proteome arity for reference. The lower triangle reports, per cell, the average value of the ratio (number of residues having pLDDT > 0.5 and predicted as disordered by `AIUPred` (score > 0.5)) / protein size.