AlphaFold predictions on whole genomes at a glance

Centre Inria at Université Côte d'Azur, France, *

November 16, 2024

Abstract

AlphaFold has opened a new era in structural biology, so that understanding the quality of its predictions is of utmost importance. Studying whole genomes, we make three contributions in this respect, with a particular focus on regions with low pLDDT values. First, we use 3D packing properties to encode a structure into a distribution amenable to comparisons using optimal transport. Summarizing this distribution using two numbers yields an informative 2D embedding, the *arity map*, which summarizes structural elements in all predictions. Second, we establish a clear correlation between the number of homologs of a protein sequence in UniProtKB, and the quality of its prediction, possibly hinting at the learning biases incurred by AlphaFold. Third, we show that certain regions of the arity map are connected with poor quality structures: proteins rich in false positive intrinsically disordered regions. The arity map thus also provides a way for detecting prediction errors that the pLDDT cannot detect.

1 Introduction

AlphaFold and AlphaFold-DB. 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. While describing the architecture of AlphaFold is beyond the scope of this paper, it may be recalled that AlphaFold uses two main modules, the Evoformer and the structure module, within which so-called transformers play a key role [2] to unveil long range couplings between amino acids. 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) [3]. 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.

Since the Protein Data Bank currently features $\sim 230K$ experimental structures against the $\sim 200M$ known protein sequences in UniProtKB, predicting structures for whole genomes is crucial [4] (SWISS-MODEL Repository). This endeavor has also been undertaken with AlphaFold with predictions deposited in the AlphaFold-DB database, see [5] and https://alphafold.ebi.ac.uk.

The goal of our work is to shed new light on essential questions connecting the AlphaFold architecture and the quality of predictions:

- Q1. Can one chart all predictions using simple and informative structural descriptors, and how do these relate to pLDDT values?
- Q2. Does the training of the AlphaFold engine entail biases in predictions ? The underlying question is naturally the balance between the information content used to train AlphaFold, and its complexity in terms of number of parameters.
- Q3. Since the challenging regions to predict are those with intrinsic disorder (IDRs), which may be found in $\sim 60\%$ of proteins, does AlphaFold process all proteins on an equal footing with respect to IDRs?

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

In the sequel, we comment previous work and position our contributions for the three questions just raised.

Q1. Structural analysis of predictions. 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 [6]. However, as indicated by a lack of association with B-factors, pLDDT does not appear to correlate with flexibility [7]. 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 [8]. In a complementary vein, the structural elements in predictions have been analyzed using the `Geometricus` algorithm [9], 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 [6]. An embedding based on shape-mers has been obtained using dimensionality reduction (t-SNE, [6, 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, it processes each structure 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 [10, 11]. Moreover, these distributions can be summarized by a small number of integers, yielding the *arity map*). In 2D, this map yields a natural embedding of all structures.

Q2. Learning biases. The `AlphaFold` pipeline is obviously very complex [1, SI]. In the context of the general bias-variance trade-off and the double descent phenomenon for large models [12] an important point is the balance between the number of parameters in `AlphaFold`, and the pieces 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![13], a simplifying process bearing similarities to coreset calculations in a more algorithmic setting learning [14].)

Such a model complexity and the commensurate information size raises the question of potential biases in the model, in particular a processing quality depending on the number of homologous sequences in `UniProtKB`. We establish a clear correlation between the number of homologs of a protein sequence in `UniProtKB`, and the quality of its prediction, possibly hinting at the learning biases incurred by `AlphaFold`.

Q3. `AlphaFold` as a predictor of 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) [15]. Indeed `AlphaFold` outperforms state-of-the-art methods such as IUPred2 on disordered region detection on a benchmark of `DisProt` targets and PDB controls [6]. 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 [16]. 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 [16].

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 Results

2.1 pLDDT statistics per organism

We start with simple analysis on pLDDT values, as these will be confronted to structural statistics in the sequel. Let us recall the following intervals for pLDDT values [1]: 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.

While the distribution of pLDDT values has been studied in selected cases, in particular for H. Sapiens in the context of conditionally folding proteins [16], we perform instead a systematic study on all genomes. We process all predictions (PDB files from `AlphaFold-DB`) of a given organism, collecting, per prediction, all/the median/the mean pLDDT value(s). These values yield three distributions, namely that for (i) all pLDDT values of all a.a. of all proteins, (ii) all median pLDDT values, and (iii) and all mean pLDDT values (Table 1 and Section S2). For the sake of conciseness, we solely discuss the case of all values.

Per genome statistics show that the fraction of a.a. with pLDDT below 0.7 lies in the range 0.078-0.408 and is rather consistent for model organisms (Table 1). However, the inspection of other organisms shows major discrepancies, with a value as low as 0.086 for P. Aeruginosa and as high as 0.584 for P. Falciparum (Table 1)

More interesting is the general shape of pLDDT distributions which, for model organisms is bimodal except for E. Coli and M. Jannaschii (Section S2). When the pLDDT values span a small range, the correlation with arity is in general low (structured protein, Fig. S26(B); unstructured protein, Fig. S28)(B). On the other hand, when the pLDDT spans a large range of values, from very low to very high, a significant Pearson correlation coefficient is generally observed (see examples Fig. S27(B), Fig. S29(B), Fig. S30(B), Fig. S31(B)).

The significant fraction of a.a. with pLDDT values inferior to 0.7 calls for further analysis to bridge the gap with structural elements, as well as the various pieces of informations used to train `AlphaFold`.

2.2 Q1. Structural studies

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. S1). 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 [10, 11]. 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 $\mathtt{arity}_{C_\alpha}(q_1) = \mathrm{CDF}^{-1}_{C_\alpha}(q_1)$ and $\mathtt{arity}_{C_\alpha}(q_2) = \mathrm{CDF}^{-1}_{C_\alpha}(q_2)$ (Fig. 1). 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. 3).

The analysis of this map provides a whole genome analysis at a glance. The x-axis $\mathtt{arity}_{C_\alpha}(q_1)$ 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. 3; 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. 3(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. 4). 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. 3(C), in red) nor in the arity plot of experimentally-confirmed disordered proteins (Fig. 6). These `AlphaFold-DB` predictions therefore represent disordered structures or inaccurate predictions.

To further characterize these proteins, we perform a hierarchical clustering of arity histograms using the Earth Mover distance as metric [10, 17]. 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. 4(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* [18]. Along the way, one maintains the connected components which appear and merge, using a Union-Find data structure [19]. 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. 1(Top, blue curve)). The same holds for unstructured proteins, excepts that the local extremum is attained at values $\sim 5 - 7$ (Fig. 1(Middle, blue curve)). The signature is more informative for proteins having both high and low quality regions (Fig. 1(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 1 *For additional examples, the reader is referred to the following figures, panel (C): folded proteins: Fig. S25, Fig. S26; unfolded proteins: Fig. S27, Fig. S28); proteins with folded and unfolded regions: Fig. S29, Fig. S30).*

pLDDT based filtration. A filtration can also be built by inserting a.a. by increasing negative pLDDT, so as to favor the emergence of connected components involving a.a. with large pLDDT values. As an example, assume that all a.a. can be segregated into (i) a first group of a.a. corresponding to an unstructured region on the N-ter side, with low confidence (say pLDDT < 0.5); and (ii) a second group of a.a. corresponding to a structured region on the C-ter side, with high confidence (say pLDDT > 0.5). Along the insertion, one observes the formation of the former and the latter regions (Section S1, Fig. S2). For a region with random pLDDT values in a range, the insertion of individual a.a. results in a well behaved curve peaking at half of the amino acids, and whose maxima if conjectured to be $1/4$ of the number of a.a. (Fig. S3; conjecture 1).

The case of `AlphaFold` predictions follows these patterns (Fig. 2, left column, top and middle cases), with a clear unimodal distribution upon selecting salient extrema. The cases of high interest are those with low and high confidence regions, as their c.c. curve again exhibits a clear bimodal profile, as evidenced by the salient extrema (Fig. 2(Bottom)). On this example, the plateau at five connected components corresponds

to five helices–four long ones plus a shorter one. The plot enables estimating the number of amino acids on the well folded regions as $\sim 781 - 4 \times 85 = 440$ – using the 1/4 conjecture mentioned above. On this example, the sum of the size of helices being of ~ 400 amino acids.

The persistence diagrams associated with the construction of the filtration also exhibit the dynamics of connected components (Fig. 2, right column). For example, for the protein from D. Melanogaster, one clearly sees the dynamics for the high (resp. low) confidence region of the molecule on the left (resp. right) part of the diagram.

Remark 2 *For additional examples, the reader is referred to the following figures, panel (E): folded proteins: Fig. S25, Fig. S26; unfolded proteins: Fig. S27, Fig. S28; proteins with folded and unfolded regions: Fig. S29, Fig. S30.*

Statistics to query whole genomes. The analysis tools just presented can naturally be used to query a whole organism so as to identify proteins with specific properties.

arity signature–for which one can also vary the number of points and their positions (Section S1, Def. 1), the connected components curves (for arity and pLDDT), and their salient points, and naturally the persistence diagrams.

2.3 Q2. Learning biases via correlations with homologous sequences

We want to test whether a correlation exists between the quality of every AF2 structural prediction and its number of homologs available on the `UniProtKB` database, on which the algorithm has been trained.

For each predicted protein of each proteome, we run a homology search over `UniProtKB` (SwissProt + TrEMBL, 2024-10-20) with `mmseqs2 easy-search`. Given the very large amount of queries, we chose `MMseqs2` over other homology search strategies (BLASTP, HMMer3) because of its time efficiency and sensitivity. Much of MMseqs rapidity comes from the prefiltering steps, which limit the number of sequence comparisons performed by the algorithm. If run with standard parameters, this strategy risks to limit the maximum number of homologs the algorithm finds per query. In order to get a reliable homolog count, we allow the maximum hits threshold parameter `-max-seqs` to attain 1,000,000. In order to further optimize the speed, we also use the parameters `-max-rejected 10` (the homolog search for a given query stops after ten consecutive rejected alignments, exploiting the sequence sorting done by the prefiltering steps) and `-min-ungapped-score 30` (restrict the rejection condition during the prefilter alignments). The E-value threshold for a match is kept to the standard `0.001`. We remind that the E-value often calculated in homology search algorithms is an estimator bound between 0 and 1 of the Karlin-Altschul statistics. For E-values < 0.001, it approximates the p-value to refuse the null hypothesis that a sequence with a score at least as high as the score of the hit found is not a homolog of the query sequence.

Counting the number of homologs for each arity signature yields a clear correlation between the arity and the number of homologs is observed (Fig. 5). (We note in passing that a similar observation has been made for selected protein [20, Fig.2].) This observation may stem from the fact that structured proteins have more homologous sequences in nature, or may be related to biases in `AlphaFold`– see also the next paragraph on IDRs. These alternatives will be studied using as reference a non redundant set of sequences, such as *e.g.* the one selected by the distillation procedure used by `AlphaFold`.

2.4 Q3. The case of 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 pLDDT threshold of 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. We expect this kind of incompleteness to be proportional to the extent of the predicted disordered regions, and not to depend on other structural characteristics of the `AlphaFold` prediction. This does not seem to be the case: according to `AlphaFold`, there is a concentration of missing IDRs for proteins having $\texttt{arity}_{C_\alpha}(0.25) < 7$ (Fig. 6). 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.

3 Outlook

This work introduces the arity map, a novel representation of structural elements in proteins, based on a filtration of amino acids depending on their 3D environment. The arity map charts the space of proteins and provides an effective tool to query specific structural properties. At the single molecule level, the persistent components singled out along the filtration can be used to compare structures. At the genome-wide level, the ground space of the arity map is useful to study specific properties. Studying the number of homologs across the arity hints at learning biases of `AlphaFold`. Studying false positives in terms of IDRs shows that `AlphaFold` has mild performances in a specific region of the arity map.

In the future, using the arity map in combination with additional parameters may provide uncanny insights onto subtle properties, including the pathogenicity of single nucleotide mutations or interactions.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

[20] Alessandra Carbone, Elodie Laine, and Gianluca Lombardi. Unlocking protein evolution insights: Efficient and interpretable mutational effect predictions with gemme. *bioRxiv*, pages 2024–08, 2024.

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

[22] 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.

4 Artwork

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**

Table 1: **Cumulated distribution function for pLDDT values of all a.a. of all proteins in a genome.** See section S2 for statistics on median and mean values. Top: model genomes. Bottom: five *global health* proteomes. Min and max values per column stressed in bold.

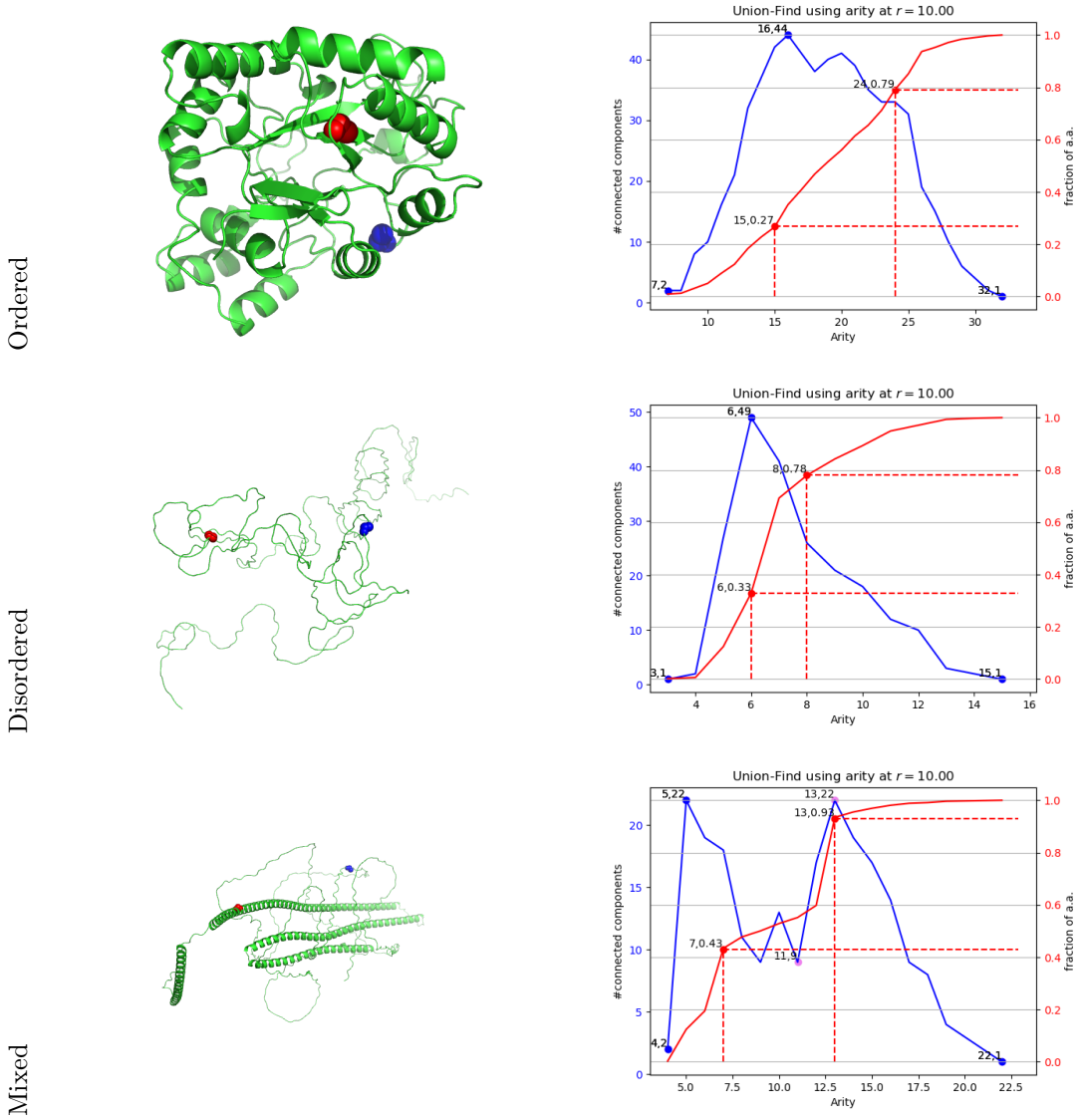
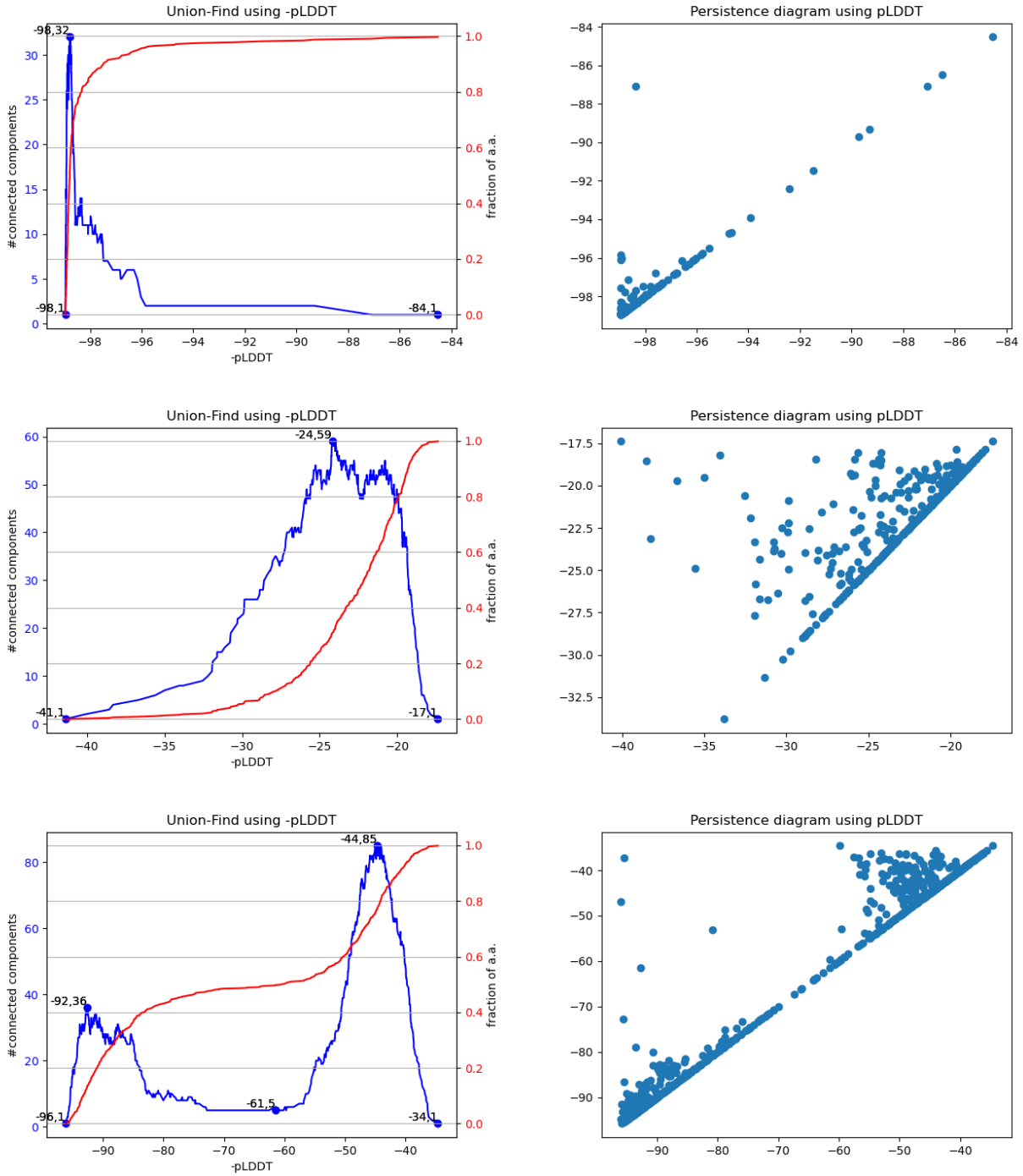


Figure 1: **Arity signature and arity based filtration on three prototypical examples.** Each structure is represented in cartoon mode; the blue (resp. red) a.a. is that featuring the lowest (resp. highest) pLDDT value. The **arity filtration** is the evolving graph defined by the connected components 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)].



Union-Find using -pLDDT

Persistence diagram using pLDDT

Figure 2: **pLDDT based filtration on the three prototypical examples of Fig. 1. (Top)** (H. Sapiens; AF-P15121-F1-model_v4; 316 amino acids.) **(Middle)** (Zebrafish; AF-A0A0G2L439-F1-model_v4; 449 amino acids.) **(Bottom)** (D. Melanogaster; AF-Q9VQS4-F1-model_v4; 781 amino acids.) 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.

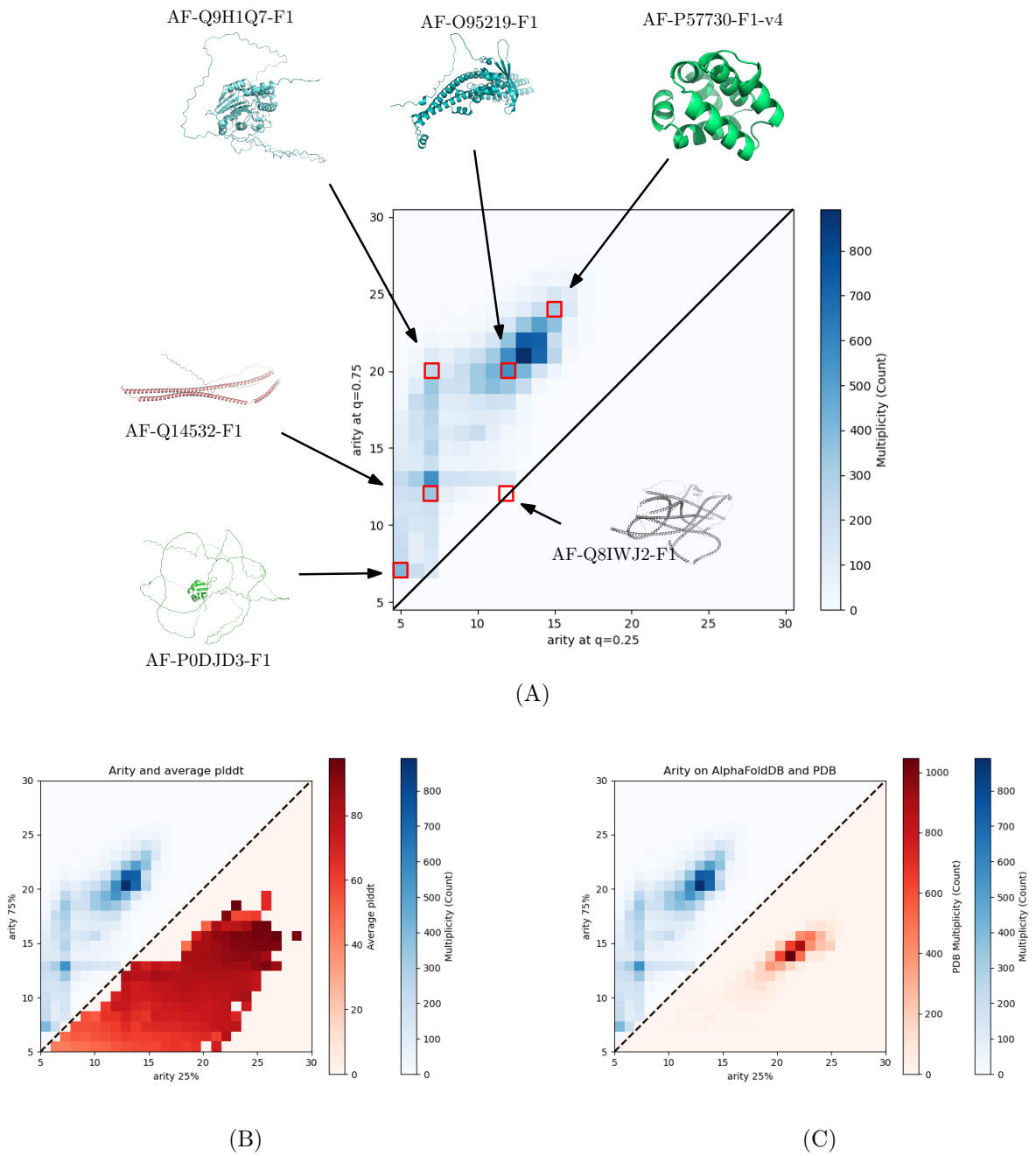


Figure 3: **H. Sapiens: arity map. (A)** Each bin corresponds to all `AlphaFold` predictions harboring a given arity signature. The upper triangle of the heatmap shows the protein count with for a prescribed arity signature. **(B)** Lower triangle: average pLDDT per bin, with mirrored coordinates. Arity 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.

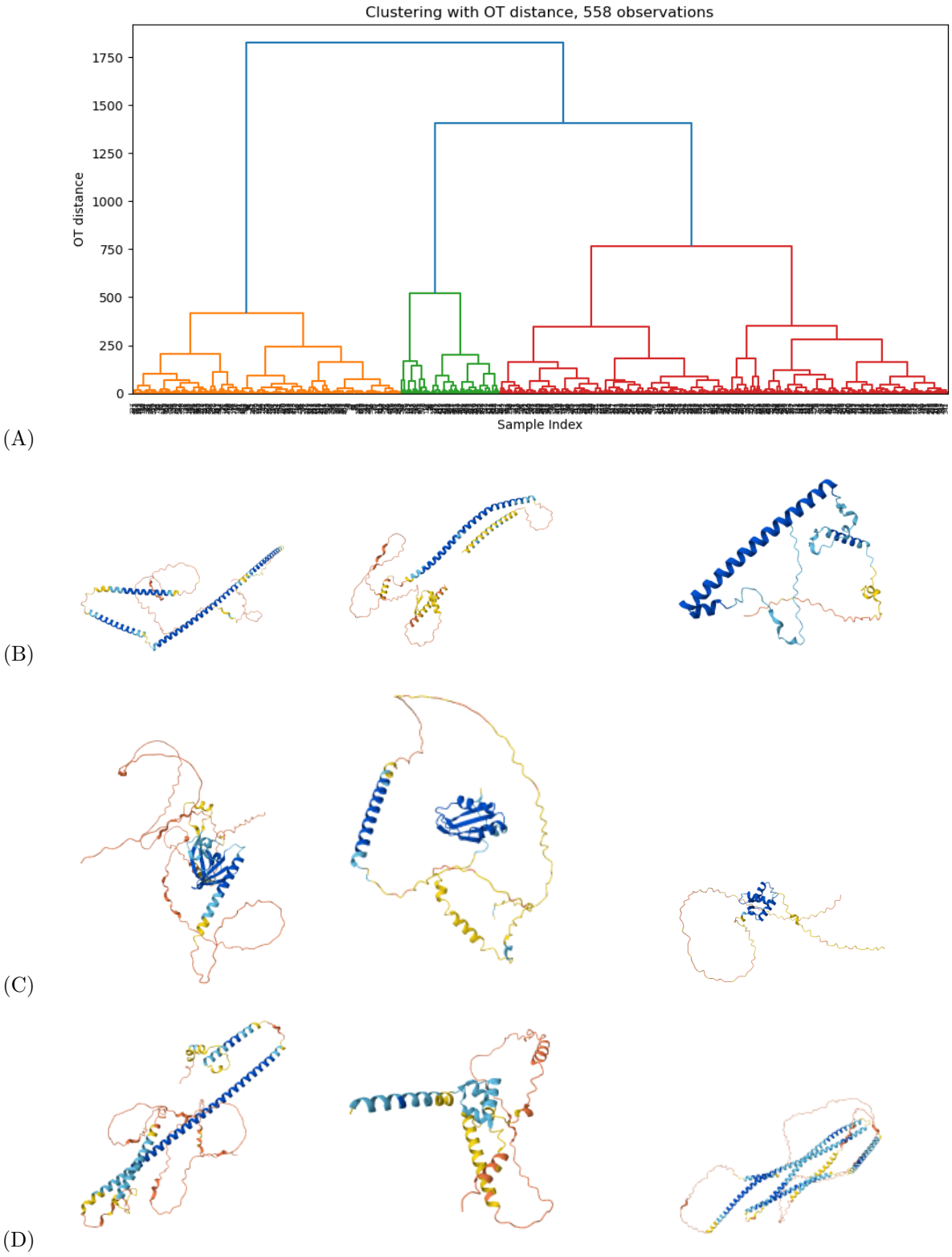


Figure 4: **H. Sapiens: hierarchical clustering of the 558 structures in the 7x13 cell of the arity map, with three random structures from every cluster. (A)** Ward hierarchical clustering of the structures using the optimal transportation distance on the normalized arity histograms H_{C_α} for each pair. **(B)** First cluster independent helices; **(C)** Second cluster: independent helices and a small β core; **(D)** Third cluster: packed helices.

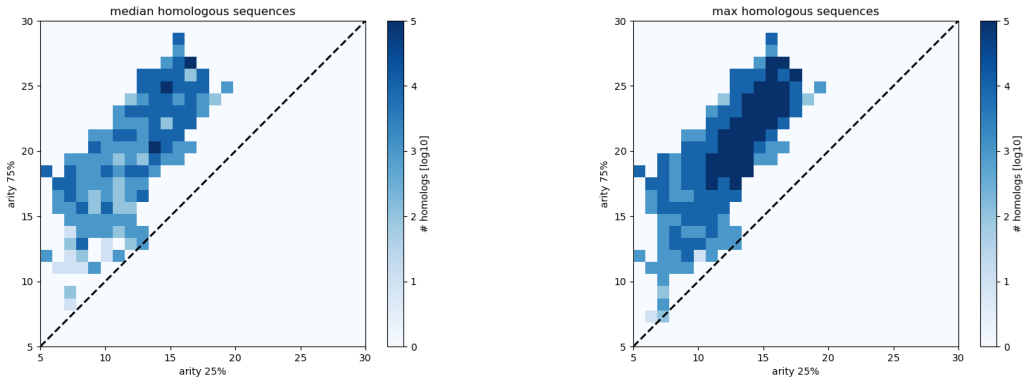


Figure 5: **H. Sapiens: the number of homologous sequences correlates with arity**. Figure **(Left)** (resp. **(Right)**) shows the median (resp. maximum) number of homologous `UniProtKB` sequences for the structures in each arity signature bin. Structured proteins (high-arity signatures, top right corner) usually find more homologous than partially or totally disordered proteins.

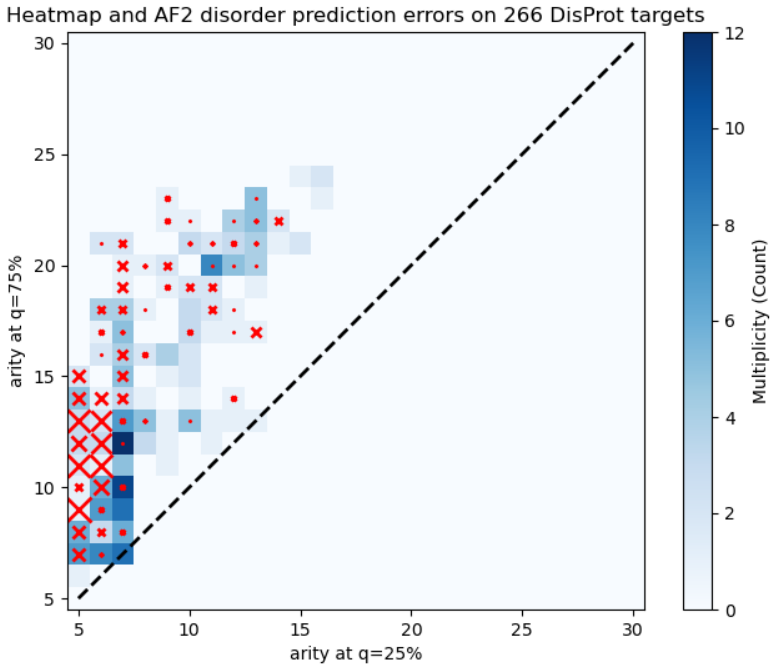


Figure 6: `DisProt` **dataset, 266 proteins lacking structural data.** The size of the red crosses over the heatmap bins tells how often a pLDDT of less than 50 does not identify an amino acid in an experimentally annotated disordered region. The leftmost region of the plot shows a high enrichment in such cases.