A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum

Apolline Bruley, Tristan Bitard-feildel, Isabelle Callebaut, Elodie Duprat

Received: 1 August 2022 | Revised: 14 October 2022 | Accepted: 18 October 2022

DOI: 10.1002/prot.26441

RESEARCH ARTICLE

PROTEINS WILEY

A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum

Apolline Bruley | Tristan Bitard-Feildel | Isabelle Callebaut | Elodie Duprat

Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Paris, France

Correspondence
Isabelle Callebaut and Elodie Duprat, Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France.
Email: isabelle.callebaut@sorbonne-universite.fr (I.C.) and elodie.duprat@sorbonne-universite.fr (E.D.)

Funding information
Agence Nationale de la Recherche, Grant/Award Numbers: ANR-17-CE12-0016, ANR-14-CE10-0021, ANR-19-CE01-0005; Institut National du Cancer, Grant/Award Number: PLBIO14-299; Muséum National d'Histoire Naturelle, Grant/Award Number: UMS2700-PCIA

Abstract

Order and disorder govern protein functions, but there is a great diversity in disorder, from regions that are—and stay—fully disordered to conditional order. This diversity is still difficult to decipher even though it is encoded in the amino acid sequences. Here, we developed an analytic Python package, named *pyHCA*, to estimate the foldability of a protein segment from the only information of its amino acid sequence and based on a measure of its density in regular secondary structures associated with hydrophobic clusters, as defined by the hydrophobic cluster analysis (HCA) approach. The tool was designed by optimizing the separation between foldable segments from databases of disorder (DisProt) and order (SCOPe [soluble domains] and OPM [transmembrane domains]). It allows to specify the ratio between order, embodied by regular secondary structures (either participating in the hydrophobic core of well-folded 3D structures or conditionally formed in intrinsically disordered regions) and disorder. We illustrated the relevance of *pyHCA* with several examples and applied it to the sequences of the proteomes of 21 species ranging from prokaryotes and archaea to unicellular and multicellular eukaryotes, for which structure models are provided in the AlphaFold protein structure database. Cases of low-confidence scores related to disorder were distinguished from those of sequences that we identified as foldable but are still excluded from accurate modeling by AlphaFold2 due to a lack of sequence homologs or to compositional biases. Overall, our approach is complementary to AlphaFold2, providing guides to map structural innovations through evolutionary processes, at proteome and gene scales.

KEYWORDS
AlphaFold protein structure database, hydrophobic cluster analysis, IDPs/IDRs, protein foldable segments, soluble and transmembrane domains

Abbreviations: aa, amino acids; AF2, AlphaFold2; AFDB, AlphaFold Protein Structure Database; CAID, Critical Assessment of protein Intrinsic Disorder; DisProt, Database of Intrinsically Disordered Proteins; FS, foldable segment; HCA, Hydrophobic Cluster Analysis; IDP/IDR, intrinsically disordered protein/region; OPM, orientations of proteins in membrane; PDB, Protein Data Bank; pLDDT, predicted local distance difference test; RSSs, regular secondary structures; SCOPe, structural classification of proteins—extended; TM, transmembrane.

Isabelle Callebaut and Elodie Duprat should be considered joint senior authors.

1 | INTRODUCTION

Protein order has been largely explored by experimental approaches so that the protein fold universe has been widely mapped.[1–3] This has led to a comprehensive inventory of the combinations according to which regular secondary structures (RSSs) are assembled to form compact, well-organized 3D structures, associated with specific functions.[4] Over the years, the structure–function paradigm has however evolved to integrate intrinsically disordered proteins/regions (IDPs/IDRs), which lack well-defined 3D structures under physiological conditions but fulfill a variety of functions, in particular in signaling and regulatory pathways.[5–8] IDPs/IDRs are characterized by a heterogeneous spatiotemporal structural organization.[9] They correspond to very diverse entities, from highly extended, heterogeneous unstructured states to compact but disordered molten globules.[10] They include short linear motifs and longer regions promoting molecular recognition and protein–protein interactions, which have led to elaborate specific classification schemes related in particular to their amino acid sequence characteristics.[6,11] IDPs/IDRs were also shown to play a key role in the formation of higher-order assemblies and in the control of many cellular processes via their participation in biomolecular condensates, through multivalent interactions leading to liquid–liquid phase separation.[12,13] IDPs/IDRs are generally defined by a heterogenous ensemble of conformations, undergoing rapid interconversion[14,15]; some can fold into a unique conformation upon binding with a partner or within oligomeric complexes,[16] while some cases of disorder maintained in the bound state were also described.[17]

Characterization of IDPs/IDRs is challenging as they are generally "unseen" by traditional structural biology methods and are therefore considered as the dark side of protein universe.[18] Their identification at large scale relies on computational methods that predict them directly from the information of the amino acid sequence.[19] Some predictors are trained on experimental annotations of protein disorder, as stored in the DisProt database,[20] while others are not, relying on physicochemical properties and predicting disorder as lack of, or deviation from order.[21,22] Using such predictive tools, IDPs/IDRs were found abundant at the proteome level, making up approximately 30% of residues in the human proteome and up to 50% in some unicellular eukaryotes such as parasitic protozoa, enriched in long IDRs (with at least 30 consecutive disordered residues).[23–26] In contrast, the fractions of disordered residues represented less than 28% in most archaeal and bacterial proteomes. The quality of disorder predictors has improved over time, in particular by considering deep learning techniques and evolutionary information, as illustrated in the recent Critical Assessment of protein Intrinsic Disorder (CAID).[27] The predicted local distance difference test (pLDDT) introduced in the recent AlphaFold2 (AF2) predictor, a deep learning program which predicts 3D structures with an unprecedented accuracy[28] and which was applied at proteome scale,[29] was also shown to provide a good metric for identifying order and disorder.[30–32]

Only a few computational approaches have addressed the issue of the predictions of different, multiple states (or flavors) of IDPs/IDRs, while not considering evolutionary information.[33,34] Here, we propose an approach for appreciating the disorder/order degree in a protein segment from the only information of its amino acid sequence, which is based on a measure of the overall density in RSSs, as predicted through Hydrophobic Cluster Analysis (HCA). HCA-based hydrophobic clusters match the positions of regular secondary structures constituting the building blocks of folded domains.[35–39] The hydrophobic alphabet and rules used for the definition of hydrophobic clusters have been supported by comparison with experimental data, and the method has been successfully applied, for instance, to the identification of remote relationships based on the conservation of 2D signatures associated with hydrophobic clusters (see Dataset S1 for a complete description of the method). The use of a simple hydrophobic/nonhydrophobic dichotomy (rather than an hydrophobicity scale), associated with the use of a two-dimensional net for defining the amino acid neighborhood, offers an efficient way to reveal these signatures in remotely related sequences.[40] We have previously developed a tool, called *SEG-HCA*, for the delineation of regions with a high density in hydrophobic clusters, which have been shown to correspond to domains which have the ability to fold, either in an autonomous way or upon contact with partners.[41,42] Contrasting with otherwise performant methods based on evolutionary couplings[43] or even earlier tools based on propensities to be in ordered or disorder states,[44], the advantage of *SEG-HCA* is to allow the prediction of foldable domains from the only information of a single amino acid sequence, without the prior knowledge of homologous sequences or consideration of pre-calculated propensities. Hence, by identifying structurally homogeneous entities (i.e., the foldable segments), this approach allows to highlight, in absence of evolutionary information, regular secondary structures that are likely to interact, however without being able to predict how they interact (which is possible by taking into account residues co-evolution). At proteome scale, *SEG-HCA* "order" predictions totalize more amino acids than the "undisordered" predictions performed by the popular IUPRED tool,[45,46] which captures the inter-residue interaction capacity by energy estimation.[42] The overlap between order and disorder detected by this comparison concentrates small sequences that are able to undergo disorder-to-order transitions.[42]

We optimized the residue weights of the here-proposed density metric for an optimal separation of foldable domains found in reference databases of order (soluble and transmembrane domains extracted from SCOPe and OPM, respectively) and disorder (DisProt), thereby deconvolving the spectrum of disorder according to the order/disorder ratio and specifying different types of disorder. Using this scoring scheme, we analyzed the per-residue confidence (pLDDT) scores of AF2 structural predictions in 21 reference proteomes, for species ranging from prokaryotes and archaea to unicellular and multicellular eukaryotes with different lifestyles.[29,47] Our analysis provides new elements to distinguish cases where low-confidence structural predictions are indeed related to disorder, as now commonly reported,[30-32] from those of domains which are foldable but whose structures cannot be accurately predicted due to AF2 intrinsic limitations. Overall, the complementarity of *pyHCA* and AF2 provides guides to map structural innovations through evolutionary processes, at proteome and gene scales.

2 | MATERIALS AND METHODS

2.1 | Delimitation of foldable segments within protein sequences

The HCA methodology is described in details in Dataset S1. *SEG-HCA*[42] was previously developed to automatically delineate regions with high density in hydrophobic clusters, constituting potential "foldable" domains within protein sequences. The new version of *SEG-HCA* was rewritten for speed and includes new functionalities, including the calculation of HCA score (see below). In addition, we rewrote and packaged in a python module the tools *SEG-HCA (in a function named segment)*, *TREMOLO-HCA* (Traveling through REMOte homoLOgy)[48] and the HCA drawing, as well as provided some utility scripts to analyze results, adding information on amino acid conservation and taxonomy. These tools can significantly help detection of hidden relationships between sequences, as repeatedly performed using the HCA approach in an expert-based way (see Ref. 41 and Dataset S1). The package can be used as a Python library, named *pyHCA*, or as a standalone tool, named *HCAtk*, and is provided at https://github.com/DarkVador-HCA/pyHCA under the CeCILL-C license agreement.

2.2 | HCA score implementation

The HCA score was introduced in order to describe the general composition of hydrophobic clusters and hydrophobic amino acids in foldable segments. Each residue of a protein sequence is associated with a class regarding the residue type and hydrophobicity. Such a residue is either (a) in a hydrophobic cluster and hydrophobic according to the common HCA alphabet, which considers as strong hydrophobic the seven amino acids V, I, L, M, F, Y, and W, (b) in a hydrophobic cluster and nonhydrophobic, or (c) outside a hydrophobic cluster. A value is assigned to each class and the HCA score (S_{HCA}) is computed as follows:

$$S_{HCA} = \sum_{i=1}^{N} w_{HCA}(seq_i, class)/N. \quad (1)$$

with $w_{HCA}(seq_i, class)$ the weight associated with the i amino acid of a sequence from a given class, and N the sequence length.

Therefore, the HCA score scales with the density of hydrophobic clusters and of strong hydrophobic residues within clusters. As the HCA score calculation motivation is to provide an estimation of the globular character, *i.e.*, the foldability of a domain, the weight of each residue was optimized within each of the three classes to minimize the overlap between the distributions of the HCA scores calculated on nonredundant disordered foldable segments from DisProt v8.0.2 sequences and nonredundant foldable segments of globular proteins from the SCOPe (soluble domains) and OPM databases (transmembrane domains), respectively (see below for details on the datasets). During the optimization step, the possible allowed values of the weights were discrete in the [−10; 10] interval ([−1; 1] scaled to one order of magnitude for better readability). For a given combination of weights, the distribution overlap (i.e., the criteria to be minimized) was estimated by histogram intersection as follows:

$$\sum_{i=1}^{B} \min(H_1(i), H_2(i)),$$

with B the number of histogram bins (set to 60), and $H1(i)$ and $H2(i)$ the values in normalized histograms of HCA scores for nonredundant dataset Disprot v8.0.2 and nonredundant dataset SCOPe and OPM for bin i, respectively. The optimization was achieved using 10-fold cross-validation, repeated 10 times. We selected the set of parameters that was most often optimal among the 10 iterations, and that allowed to achieve the least overlap between disorder and order.

This implementation is an updated version (v2) of a previous one, described in Ref. 49 and applied in investigations such that reported in Ref. 50, where the disorder dataset corresponded to sequence segments predicted by Mobi-DB[51] on the DisProt v7.0 sequences. The *pyHCA* GitHub repository has been updated accordingly (https://github.com/DarkVador-HCA/pyHCA).

2.3 | Sequence datasets

2.3.1 | Sequence redundancy filters

For each dataset described below (b–e), sequence redundancy was addressed using the *MMseqs2*[52] clustering module (clustering mode 1, sensitivity 8) with a sequence identity threshold of 30% and a coverage threshold of 90%. The nonredundant sets of sequences (Dataset S2) comprised the representative sequences of each cluster.

2.3.2 | Disordered segments

Disordered sequences, as assessed from experiments and manually curated, were extracted from the reference database DisProt v8.0.2 (8 254 sequences,[20]; https://disprot.org/). The corresponding nonredundant set DisProt comprises 3 166 sequences.

2.3.3 | Soluble domains

27 543 sequences of soluble domains with known 3D structures were collected from the Structural Classification of Proteins—extended (SCOPe) v2.0.7 database ([53]; https://scop.berkeley.edu/) as provided by Astral repository with 95% identity filter. The SCOPe classification of these entries according to their content in regular secondary structures was as follows: 4 974 all-alpha domains (class a), 7 622 all-beta domains (class b), 8 250 alpha/beta domains (class c), 6 697 alpha +beta domains (class d). Our nonredundant dataset SCOPe comprises 10 885 domain sequences with the SCOPe a-d classes represented by 2 507, 2 511, 2 841 and 3 026 entries, respectively.

2.3.4 | Transmembrane domains

The orientations of proteins in membranes (OPM) database ([54]; https://opm.phar.umich.edu/) was evidenced to include the largest number of membrane proteins with known 3D structures.[55] The OPM entries annotated as transmembrane domains were downloaded on August 30, 2021 and include 3 classes: alpha-helical polytopic or multi-pass (140 superfamilies, 5 381 entries), bitopic or single-pass (69 superfamilies, 1 151 entries) and beta-barrels transmembrane (35 superfamilies, 601 entries) domains. 35 sequences only formed by unknown residues (replaced in OPM by alanines) were omitted. We performed a first redundancy treatment to remove repeated sequences (using *MMseqs2* clustering module with a sequence identity threshold of 95% and keeping the longest sequence of each cluster), leaving only 3 051 unique TM domains in total. For each domain, OPM provides the information of the calculated transmembrane (TM) segment boundaries. In order to delineate the whole membrane-spanning domains, we included loops with lengths less than or equal to 30 residues in our sequence dataset. According to the distribution of loop lengths in TM domains and to the minimum size of known globular domains, TM segments separated by more than 30 residues are likely to encompass nested soluble domains.[56] In order not to include these cases of large loops in our sequence dataset, we only kept the first 15 residues after the first TM segment and the last 15 before the second TM segment. If an extended segment boundary falls inside a hydrophobic cluster, we moved it to include the whole hydrophobic cluster and the 4 following residues (i.e., the minimal distance considered to separate contiguous hydrophobic clusters, see Dataset S1). This pruning needed to be applied to 1 882 sequences (see Dataset S2 for details). The OPM nonredundant dataset comprises 1 698 sequences: 1 330, 165, and 203 annotated as alpha-helical polytopic, bitopic and beta-barrels transmembrane domains, respectively.

2.3.5 | Proteomes from AlphaFold Protein Structure database v1

The amino acid sequences, the 3D structure predictions and the corresponding per-residue model confidence values (pLDDT) were downloaded from the AlphaFold Protein Structure database (AFDB) v1[47] (https://alphafold.ebi.ac.uk, downloaded on July 21, 2021) for the reference proteomes of 21 model organisms. pLDDT values estimate on a per residue basis how well the predicted structure would agree with the experimental 3D structure and is scaled between 0 and 100 as follows: very low (pLDDT ≤ 50), low (50 < pLDDT ≤ 70), confident (70 < pLDDT ≤ 90), very high (pLDDT > 90).

2.4 | Figure creation and statistical analyses

3D structures were visualized with the UCSF Chimera package.[57] Statistical analyses were performed using the R software, version 4.1.2 ([58]) and the Python Language (http://www.python.org), versions 3.7.6 (for the HCA score implementation and AFDB v1 analyses) and 3.6.3 (for the classification of order and disorder from DisProt, SCOPe and OPM databases). Standardized principal component analysis (PCA) and hierarchical clustering on principal components were performed using the R package Factoshiny, version 2.4 (https://cran.r-project.org/web/packages/Factoshiny). Mean comparisons by nonparametric Mann–Whitney *U* test were performed using the Python scipy library, version 1.4.1 (https://scipy.org). Graphics were generated using the Python libraries matplotlib, version 3.0.3 (https://matplotlib.org) and seaborn, version 0.11.2 (https://seaborn.pydata.org).

3 | RESULTS

3.1 | Exploring the degree of order and disorder in reference databases using HCA score

HCA score was introduced in order to provide a global estimation of the density of a given amino acid sequence in hydrophobic clusters and hydrophobic amino acids within hydrophobic clusters, as a proxy for order/disorder ratio in protein segments. Hence, we focused on foldable segments (as defined by the *segment* function of the *pyHCA* package) of the reference datasets and optimized the weights of this metric to best separate the different categories according to structural features.

We optimized the weights of this metric to best separate order and disorder relative to two datasets: (a) a set of 16 489 nonredundant foldable segment sequences from globular proteins from the reference databases SCOPe for soluble domains (14 624 segments) and OPM for transmembrane domains (1 865 segments) and (b) a set of 3 276 nonredundant foldable segment sequences from disordered proteins from the reference DisProt database (v8.0.2) (Dataset S3). The optimal HCA score was reached for a minimum overlap of 49% between the score distributions of the globular and disordered sequences, respectively, and with the weights of 9 for strong hydrophobic residues in hydrophobic clusters, 7 for nonstrong hydrophobic residues in hydrophobic clusters and −10 for residues outside of hydrophobic clusters (see Eq. (1) in Section 2).

A vast majority of amino acids of the SCOPe and OPM nonredundant datasets are included in foldable segments (90% and 96%, respectively) (Table 1), while non foldable segments in these datasets mostly correspond to large, hydrophilic loops (see for instance the example shown in Figure S1a). In contrast, only 50% of amino acids in the DisProt v8.0.2 dataset are included in foldable segments (Table 1), the remaining ones (i.e., in nonfoldable segments) can be thus considered as fully disordered (Figure S1b). DisProt segments can thus be analyzed after being separated into two distinct categories (i.e., foldable and nonfoldable segments). A few instances of poor coverage of folded domains of the SCOPe database by foldable segments are observed. These correspond to sequences rich in alanine (class a) and threonine/serine (class b), which are not included in the hydrophobic alphabet used for hydrophobic cluster definition, but

TABLE 1 Foldable segments (FS) in the nonredundant sequence datasets

Dataset	Class	Sequences	aa	Sequences with FS	Sequences without FS	aa in FS	aa outside of FS
SCOPe	a	2 507	347 268	2 507 (100.0%)	0 (0.0%)	309 100 (89.0%)	38 168 (11.0%)
SCOPe	b	2 511	366 584	2 511 (100.0%)	0 (0.0%)	322 970 (88.1%)	43 614 (11.9%)
SCOPe	c	2 841	688 643	2 841 (100.0%)	0 (0.0%)	633 955 (92.1%)	54 688 (7.9%)
SCOPe	d	3 026	455 400	3 026 (100.0%)	0 (0.0%)	410 347 (90.1%)	45 053 (9.9%)
DisProt v8.0.2		3 166	209 812	2 418 (76.4%)	748 (23.6%)	105 390 (50.2%)	104 422 (49.8%)
OPM	polytopic	1 330	186 220	1 328 (99.8%)	2 (0.2%)	181 396 (97.4%)	4 824 (2.6%)
OPM	bitopic	203	4 390	202 (99.5%)	1 (0.5%)	4 025 (91.7%)	365 (8.3%)
OPM	beta	165	37 320	164 (99.4%)	1 (0.6%)	33 986 (91.1%)	3 334 (8.9%)

Note: Total number of sequences and amino acids (aa) of each dataset, as well as their number (and percentage) of sequences with and without foldable segment(s), and the number (and percentage) of residues found in or outside of foldable segments. The datasets are those of soluble domains with known 3D structures (SCOPe), transmembrane domains with known 3D structure (OPM) and disordered segments (DisProt v8.0.) (see *Materials and Methods* for details). OPM classes have been shortened to polytopic for alpha-helical polytopic domains, bitopic for alpha-helical bitopic domains and beta for beta-barrels.

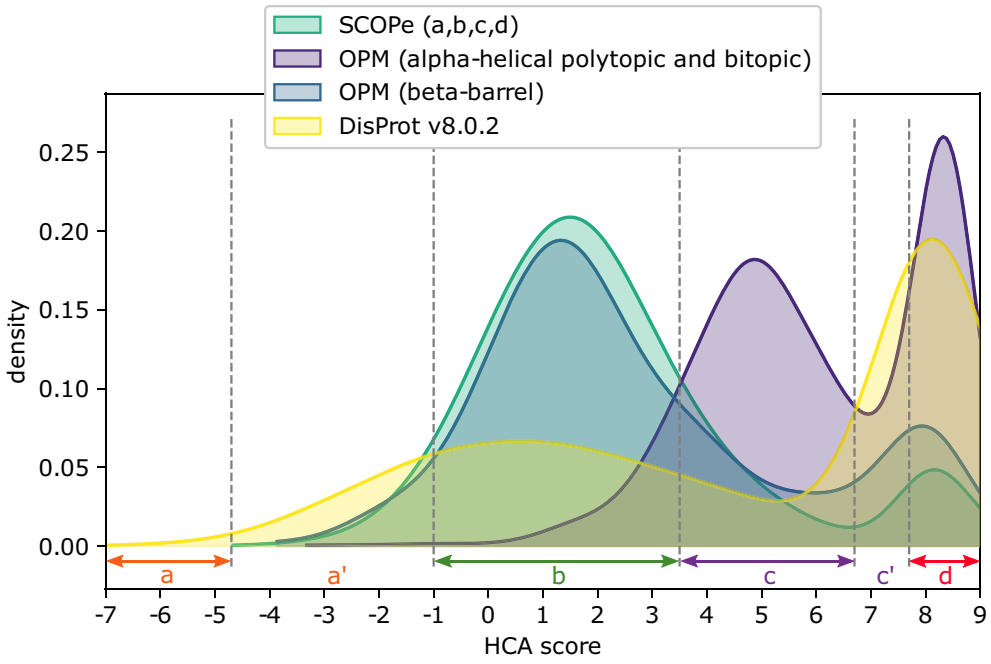

FIGURE 1 Distribution of the HCA scores calculated for the foldable segments from disordered protein regions (DisProt) and ordered protein domains (soluble domains from SCOPe and membrane domains from OPM). The considered nonredundant datasets are: DisProt v8.0.2 (HCA scores ranging from −6.95 to 9 and peaking at 0.5 and 8.1), SCOPe (HCA scores ranging from −4.67 to 9 and peaking at 1.4 and 8.1), OPM alpha-helical polytopic and bitopic categories (HCA scores ranging from −3.32 and 0.97 to 9 and peaking at 4.9 and 8.3, respectively) and OPM beta-barrel category (HCA scores ranging from −3.86 to 9 and peaking at 1.2 and 7.9). Several thresholds were fixed from these distributions, allowing to define four main classes of foldable segments (intervals a-d). The threshold of 3.5 better discriminates the HCA scores of SCOPe foldable segments from those of alpha-helical OPM foldable segments. −1 is the threshold above which are found 95% of the HCA scores computed for foldable segments from SCOPe protein domains. (a) Below −4.7: only foldable segments from disordered regions, (b) from −1 to 3.5, globular soluble segments (SCOPe) or membrane beta-barrels (OPM), (c) from 3.5 to 6.7, segments from alpha-helical transmembrane domains, (d) above 7.6, foldable segments composed of only one hydrophobic cluster. Two intermediate regions were also defined: (a′) between −4.7 and −1, in this range 62% of segments are from SCOPe and 36% from DisProt, they represent only 5% of SCOPe and 11% of DisProt; (c′) between 6.7 and 7.6, foldable segments with a dense composition in hydrophobic clusters and of short length (96% are shorter than 100 aa).

participate in the hydrophobic core (Figure S1c,d). Some other cases of low coverage actually correspond to domains stabilized by a ligand or for which the folding is cooperative or dependent of an oligomeric organization (obligate oligomers) (Figure S1e).

The HCA score values range between −7 and 9 (Figure 1, Dataset S3). Soluble domains have HCA scores included mostly in the [−1,3.5] interval (labeled b). As illustrated for the SCOPe b class, the lowest and highest values are mainly associated with the presence

and absence of large loops (thus large intercluster linkers), respectively (Figure 2A). HCA scores for the beta-barrel transmembrane class are close to those of this SCOPe distribution, while those for the alpha-helical polytopic class are higher ([3.5,6.7] interval, labeled c) (Figure 1). This is consistent with the general properties of these two classes of proteins; the beta-barrel strands, albeit longer, are indeed characterized by a periodicity of 2 in hydrophobic amino acids, as for beta-strands from globular domains, while transmembrane alpha-helices from alpha-helical polytopic segments are associated with larger numbers of strong hydrophobic amino acids dotted with charged/polar amino acids, as well as tiny ones (Gly, Ala), ensuring tight packing (Figure 2B). Beta-barrel transmembrane sequences can be distinguished from soluble globular domains by their amino acid composition and hydrophobic cluster characteristics (Figure 2B, also see Discussion). Finally, a separate category exists for all reference databases, with HCA scores above a value of 7.6 (labeled d), including

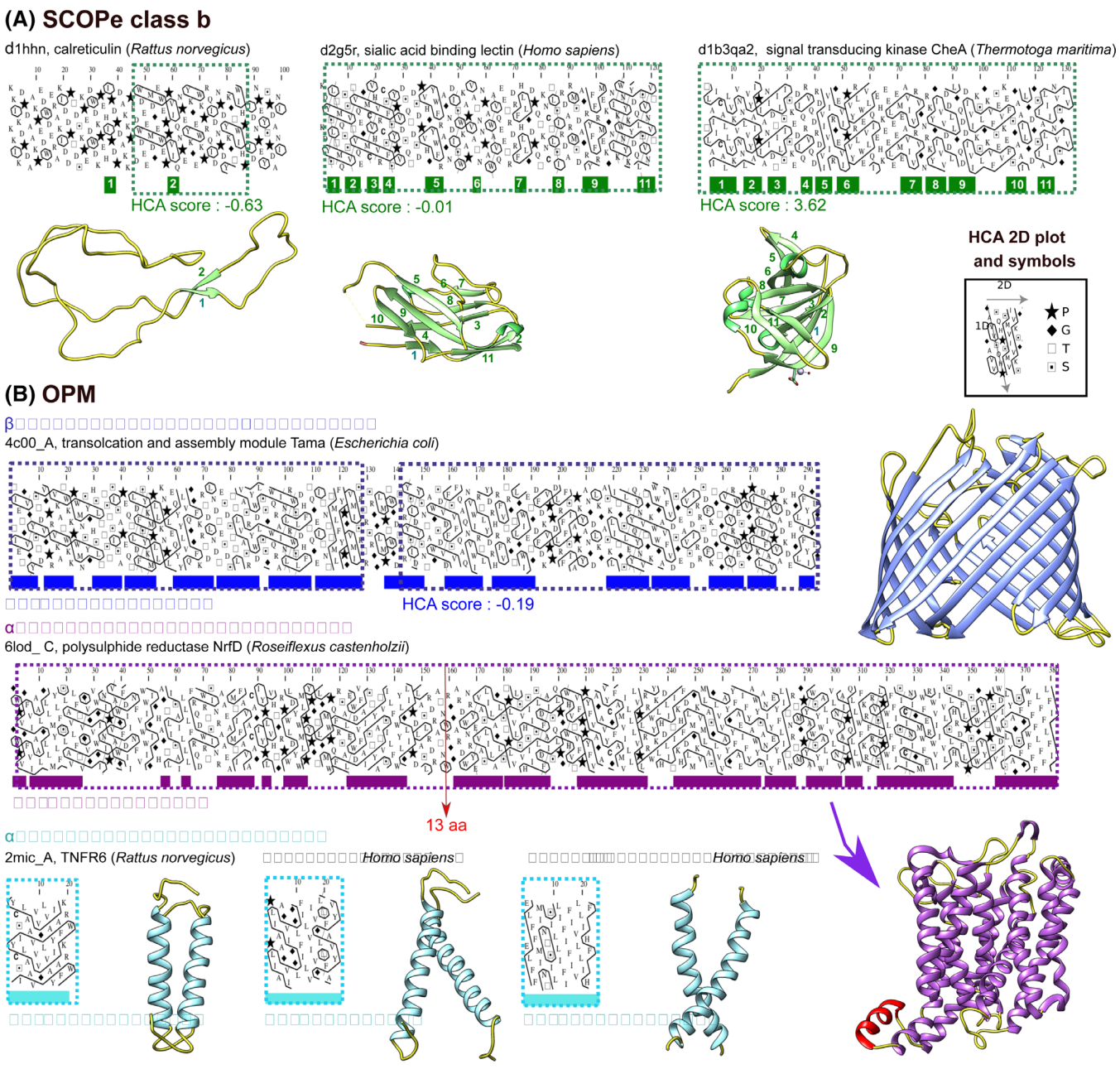

FIGURE 2 Foldable segments extracted from the SCOPe and OPM databases: HCA scores, HCA 2D plots and experimental 3D structures. The HCA plots of the sequences extracted from the SCOPe (all beta) (A) and OPM (B) databases are shown, with their foldable segments boxed (dashed lines). The corresponding HCA scores are reported below the boxes. Special symbols used in the HCA representation and the way to read the sequences and regular secondary structures (RSSs) are indicated in the inset. Positions of the RSSs, as experimentally observed from the corresponding 3D structures (ribbon representations), are reported in color below the HCA plots. The position of the 13 aa-long segment which has been removed for the OPM protein sequence in order to only keep membrane domains is highlighted in red (see Section 2 for details)

mostly short sequences covered by a single hydrophobic cluster. These correspond entirely to the foldable segments of OPM bitopic sequences (made of only one transmembrane segment, rich in strong hydrophobic amino acids, Figure 2B light blue) or are part of OPM polytopic or SCOPe soluble sequences, which have been fragmented into several parts due to the presence of large loops (see the example in Figure S1a). Overall, the HCA score represents a metric that allows to globally assess the order/disorder ratio in a single protein sequence and to unravel its main structural features (Table 1).

The HCA score distribution of foldable segments from the DisProt database is wide, ranging from −7 to 9, due to the structural heterogeneity of the corresponding sequences, reflecting diverse order/disorder ratio. Some foldable segments have low HCA score values, having hydrophobic cluster densities below those observed for folded domains, which reflect their propensity to fluctuate between disorder and order and/or to interact with partners (labeled a and a′ in Figure 1). This is for instance the case of a foldable segment from the yeast nuclear pore complex NUP2 protein, which contains FxFG repeats (shown with a red circle in Figure 3A, HCA score of −5.36 [interval a in Figure 1]). FxFG repeats are known to bind transport factors of the importin-beta/karyopherin-beta family, which function as carriers for many nuclear trafficking processes.[59] They bind in hydrophobic pockets displayed at the surface of HEAT repeats, in which the phenylalalanine side chains are buried (modeled on Figure 3A, on the basis of a complex of FxFG peptides with importin-beta). A second example is related to foldable segments of the mouse nucleoporin ELYS, interacting with chromatin[60] (Figure 3B, no 3D structure available, HCA score of −2.84 [interval a′ in Figure 1]). A third example is that of a cadherin 1 segment, whose 3D structure has been partially solved in complex with catenin beta 1[61] (Figure 3C, HCA score of −1.62 [interval a′ in Figure 1]). Some other DisProt foldable segments are falling into the "folded, soluble domain" category (interval b in Figure 1), being characterized by hydrophobic clusters ratio typical of this kind of stable 3D structures. This is for instance the case of a foldable segment of the 7SK Sn RNA methylphosphate capping enzyme, which is partially disordered and undergoes a disorder-to-order conformational change upon RNA binding[62] (Figure 3D, HCA score of 0.40). Another example of a DisProt sequence, totally covered by a foldable segment, is shown in Figure S2a, with a cluster composition and density similar to that observed in globular domains, albeit with a slightly low content (30%) in hydrophobic amino acids. This example corresponds to the yeast proteasome maturation factor Ump1, which is disordered when free in solution, as observed using various experimental techniques.[63,64] It however forms well-ordered secondary structures in complex with the 20 S core particle, playing a key role in the dynamical assembly of proteasome.[65] An additional example of stabilization of 3D structures in complexes is provided in Figure S2b. It is also in this interval b of HCA score, typical of well folded 3D structures, that are found molten globules (IDP0:00077), which are compact, with native-like secondary structures but disordered tertiary structures.[10] We illustrate this case with the well-known nuclear coactivator binding domain (NCBD) of the mouse CREB-binding protein (DisProt DP00348r018, aa 2059–2117; foldable segment aa 2068–2102, HCA score = −0.09),[66] which moreover folds into two different conformations depending on the binding partner[67] (Figure S2c).

Next, higher HCA scores values, typical of polytopic membrane domains (interval c in Figure 1), are also encountered in DisProt sequences. In a general way, these segments have long hydrophobic clusters (which are common in membrane domains), but form long helical structures involved in oligomeric coiled-coils. This is the case for instance of the HR2 domain of SARS-CoV-2 spike glycoprotein, which form an elongated six-helix bundle together with the HR1 domain[68] (Figure 3E, HCA score of 5.66). Another example is the flexible C-terminal part of the *Escherichia coli* antitoxin ParD, which is involved in the binding and neutralization of the ParE toxin[69] and forms upon binding long helices docking into the groove of the partner[70] (DP0033r012, HCA score = 5.15, 27 amino acids).

Finally, HCA score values above 7.6 (interval d in Figure 1) include foldable segments with a single hydrophobic cluster, as for similar segments of the SCOPe and OPM databases. Such segments frequently correspond to preformed secondary structures (MORFs) or short linear motifs (SLIMs), which fold upon binding, as illustrated here with the short linear motif of the *C. elegans* EGL-1, whose binding to CED-9 initiate cell programmed death[71] (Figure 3F, HCA score of 8.23).

Overall, applied to disordered sequences extracted from the DisProt database, the HCA score also allows to globally assess the order/disorder ratio and unravel the wide diversity (or different flavors) of disorder.

3.2 | Leveraging AlphaFold2 predictions with *pyHCA* package

The recent development of the AlphaFold2 (AF2) deep-learning approach was a huge step forward in the high-throughput prediction of 3D structures for folded regions, and was even announced as a disorder prediction tool, as regions with very low confidence largely overlap with IDRs.[30–32,72] We wanted here to compare the AF2 predictions to those of foldability and structural states, which can be made using *pyHCA*. Each proteome processed by AF2, provided by the AlphaFold Protein Structure Database (AFDB,[47] https://alphafold.ebi.ac.uk), can be described in terms of foldability and order/disorder content by the *pyHCA* package: the *segment* function allows first to quantify the proteome coverage in foldable segments, then to compute the HCA score on foldable segments, leading to their assignment to a structural class related to their position within the order–disorder continuum, as inferred in Figure 1. The link between these two metrics (one, binary, and the other, discretized from a continuum) and the per-residue metrics of uncertainty (pLDDT) defined by AF2 was here explored. Four main classes of pLDDT values, reflecting the confidence in the AF2 structural predictions, are generally considered: very high (pLDDT > 90), high (90 ≥ pLDDT > 70), low (70 ≥ pLDDT > 50) and very low (pLDDT ≤ 50).

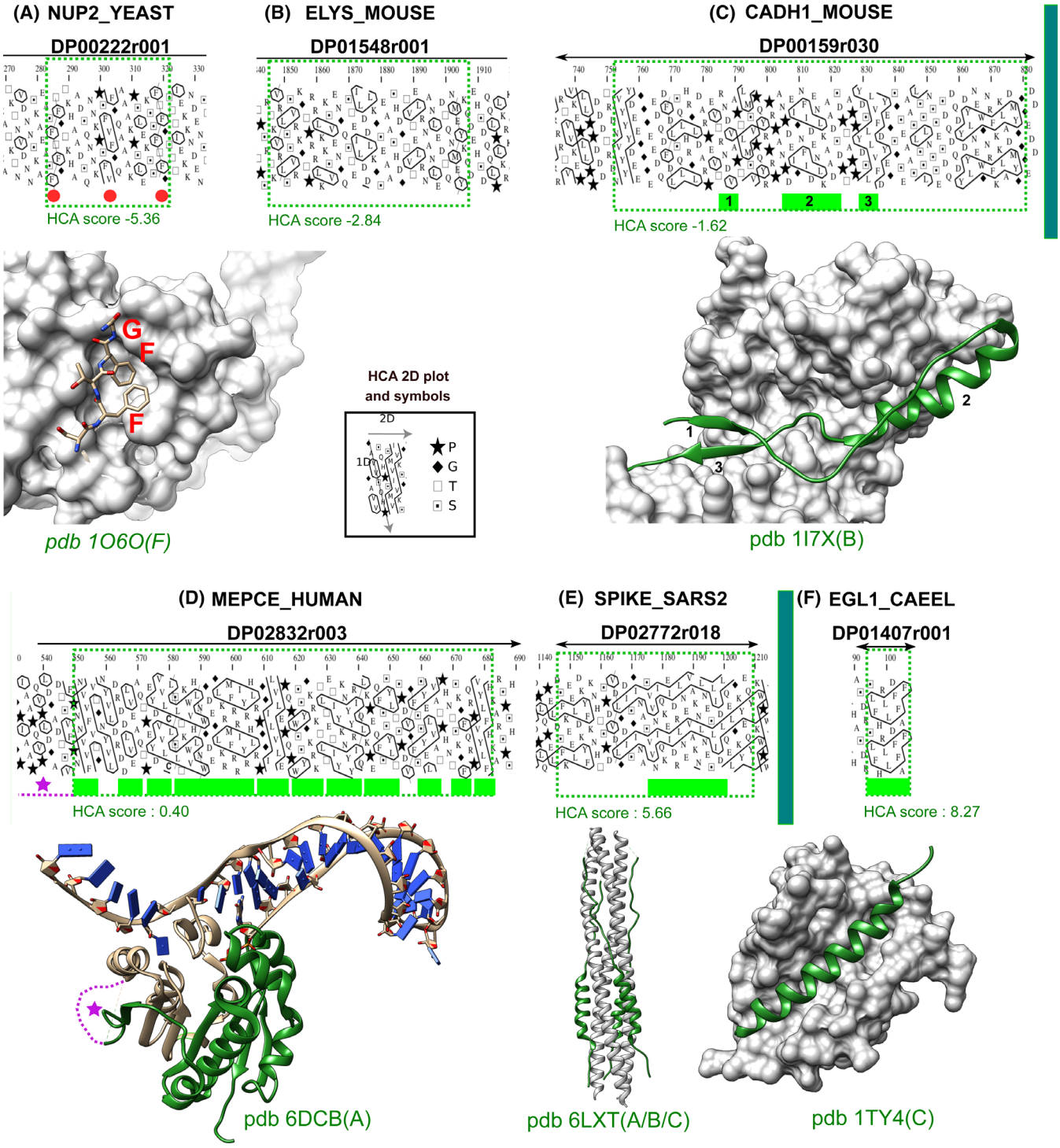

FIGURE 3 Foldable segments extracted from the DisProt database: HCA scores, HCA 2D plots and experimental 3D structures. The HCA plots of the sequences extracted from the DisProt database are shown (arrows indicate the N- or C-terminal limits of the DisProt segments). The foldable segments are boxed (dashed lines). The corresponding HCA scores are reported below the boxes. Special symbols used in the HCA representation and the way to read the sequences and regular secondary structures (RSSs) are indicated in the inset. 3D structures (in green, with their pdb identifiers) have been solved for only a few of the DisProt sequences (moreover often limited to only a part of the regions), stabilized by interaction with a partner (gray surfaces for two of them). One of the FXFG repeats of yeast NUP2 (red circles) has been modeled based on the experimental 3D structure of such a repeat in complex with importin beta-1 (*pdb 1O6O*)

We first analyzed the confidence in the AF2 structural predictions for residues in versus outside of the foldable segments, for the 362 094 sequences of the 21 reference proteomes from AFDB v1. Among this whole dataset, 81.9% of the residues are part of foldable segments (ranging from 73.1% to 96.3% for the proteomes of the parasitic protozoa *Leishmania infantum* and the autotrophic

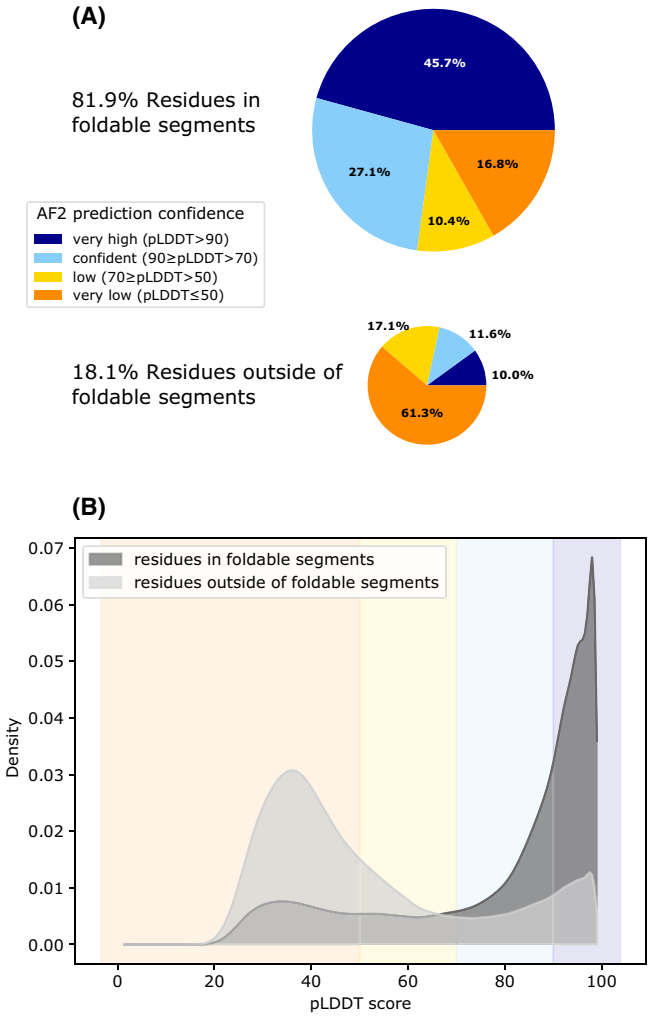

FIGURE 4 Distribution of AlphaFold2 per-residue prediction confidence scores (pLDDT) within and outside of foldable segments. (A) Distribution of residues from the AFDB 21 proteomes within foldable segments (top, 81.9% of the total) and outside of foldable segments (bottom, 18.1% of the total) in the different categories of AF2 prediction confidence. (B) Distribution of pLDDT scores for residues within (dark gray) and outside of (light gray) foldable segments. The AF2 prediction confidence categories are highlighted following the same color code as in (A)

hyperthermophilic archaeon *Methanocaldococcus jannaschii*, respectively), corresponding mostly to confident predictions: 45.7% and 27.1% residues with very high and high pLDDT, respectively (Figure 4A). Instead, the residues located outside of foldable segments (hereafter described as nonfoldable segments, representing 18.1% residues of the whole dataset) correspond mostly to low confident predictions (61.3% and 17.1% residues with very low and low pLDDT, respectively, Figure 4A). These trends are also observed for each organism separately, at the exception of *Plasmodium falciparum* whose proteome is dominated by low confident predictions, even in the foldable segments (39.7% and 13.1% residues with very low and low pLDDT, respectively) (Figure S3 and Table S1). The four prokaryotic proteomes are instead largely dominated by high confident predictions, in foldable (>72% residues with very high pLDDT) but also in nonfoldable segments (<40% residues with very low or low pLDDT). Overall, low and high pLDDT scores are thus mostly observed for amino acids in nonfoldable (full disorder) and foldable segments, respectively (Figure 4B, Table S1), thereby independently supporting the observation that AF2 low confidence predictions are significantly enriched in intrinsically disordered regions.[30,31] This is consistent with the recently reported distributions of pLDDT scores for the per-residue predictions of order and disorder.[72,73]

In order to further explore the confidence of AF2 predictions for foldable and nonfoldable segments separately, we considered regions of contiguous residues (2 aa as minimum length) within these segments where the residues are all affiliated to the same class of pLDDT values (4 classes, see above). In the nonfoldable segments, 96.1% residues with a very high pLDDT are located in such regions of homogeneous prediction confidence (3.9% residues with a very high pLDDT in nonfoldable segments are thus isolated within regions with a lower prediction confidence).

The sequence length distribution of these regions of very high prediction confidence in nonfoldable segments is illustrated in Figure S4 (only 13 sequences with length greater than 100 amino acids are observed, corresponding to 6 different proteomes). These regions fall into two distinct categories, corresponding to: (a) large loops within folded domains (Figure S5a) or linkers between folded domains (Figure S5b), which are generally disordered or flexible in isolated proteins but conditionally folded in presence of specific interactions, (b) regions which are well folded but independent of the presence of strong hydrophobic amino acids (Figure S5c–f). These sequences correspond to either (a) long coiled-coils, which form extended rod-like structures and are rich in alanine (Figure S5d) or made of acidic and basic-rich heptad repeats (Figure S5c), and to various structural repeats (left-handed parallel beta helix repeats (Figure S5e), tetratricopeptide repeats, armadillo repeats, …) or (b) domains with ion-dependent folding (calcium [Figure S5f], zinc, …). Nonfoldable sequences which form well-stable structures thus correspond to particular cases in which the fold is not conditioned by the presence of a core of strong hydrophobic amino acids, but which possess clear, distinctive amino acid composition. Conversely, as discussed in Ref. 73 foldable segments do include sequences with disorder potential, for which AF2 prediction however capture some structural features formed upon interaction (some of which already depicted in experimental structures).

The methodology developed here allows us to put the AF2 prediction made at the residue level into the context of the foldable segment to which it belongs and of its position in the order/disorder continuum. We thus calculated the distribution of residues in each pLDDT category as a function of the HCA scores of the foldable segments in which they are included (Dataset S4). Overall, we observed differences in the relative proportions of each pLDDT category as regards to the structural states (labeled a to d), previously defined based on the HCA score (Figure 5A). The highest proportion of residues with very high pLDDT (51.8%) is observed for the foldable segments with an HCA score within]1;2] interval, corresponding to the

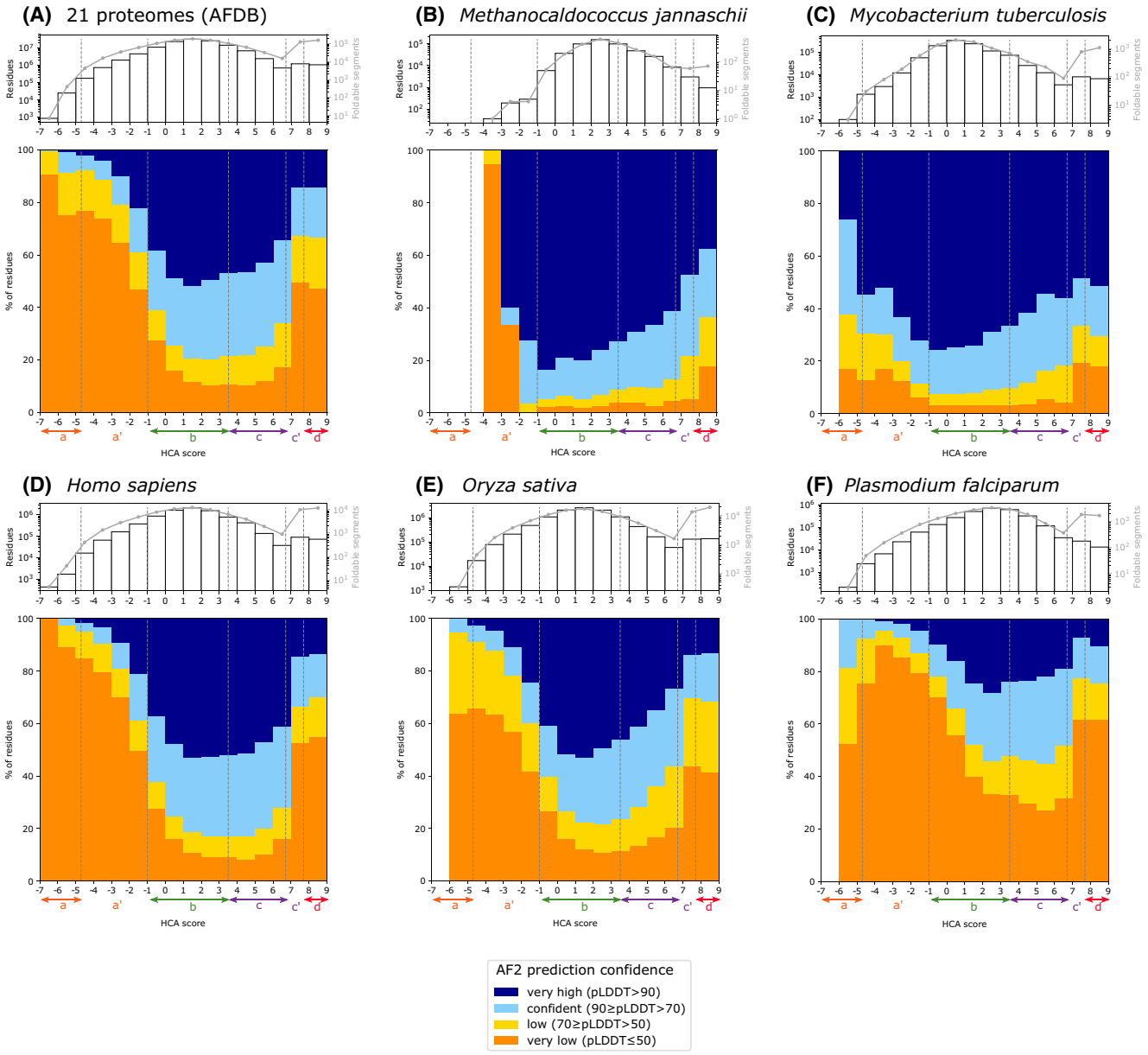

FIGURE 5 Relationship between HCA score and AlphaFold2 per-residue prediction confidence score for foldable segments. (A) AFDB v1 21 proteomes, (B–F) 5 representative proteomes from AFDB v1 (see Figure 6 and Figure S6 for details). Top: Representation of the number of residues by the barplot (axis on the left, base 10 logarithmic scale) and number of foldable segments by the gray line and points (axis on the right, base 10 logarithmic scale) belonging to foldable segments with HCA scores in a given interval. Bottom: percentage of residues in each AF2 prediction confidence categories for each HCA score range. Below the barplot are represented the foldable segment types that are likely to be found according to their HCA score, as defined in Figure 1: (a) disorder only, (a′) mainly disorder, (b) globular domains/ membrane domains (beta-barrel), (c) membrane domains (alpha-helical), (d′) mainly foldable segments with one hydrophobic cluster, (d) foldable segments with one hydrophobic cluster

typical score of soluble domains (interval b, as defined in Figure 1). Moreover, the residues with very high pLDDT represent 48.3% residues included in foldable segments with a HCA score in]−1;6], corresponding to the soluble/beta-barrel and alpha-transmembrane domains (intervas b and c). Instead, residues with very low pLDDT dominate the likely disordered foldable segments with a HCA score in]−7;−4] (74.8%), but also those with a HCA score in]+7;+9] (63.7%) corresponding to local maxima in the HCA score distribution for DisProt disordered sequences (see Figure 1). We should notice the uniformity of the length of the foldable segments with HCA score lower than +7, with a median length of 84 aa, as evidenced by the tight correlation between the number of residues and the number of foldable segments within the corresponding intervals of HCA score (see top graph in Figure 5A, with black bars and gray line, respectively). Instead, the foldable segments with HCA score higher than +7 are much shorter, with a median length of 6 aa.

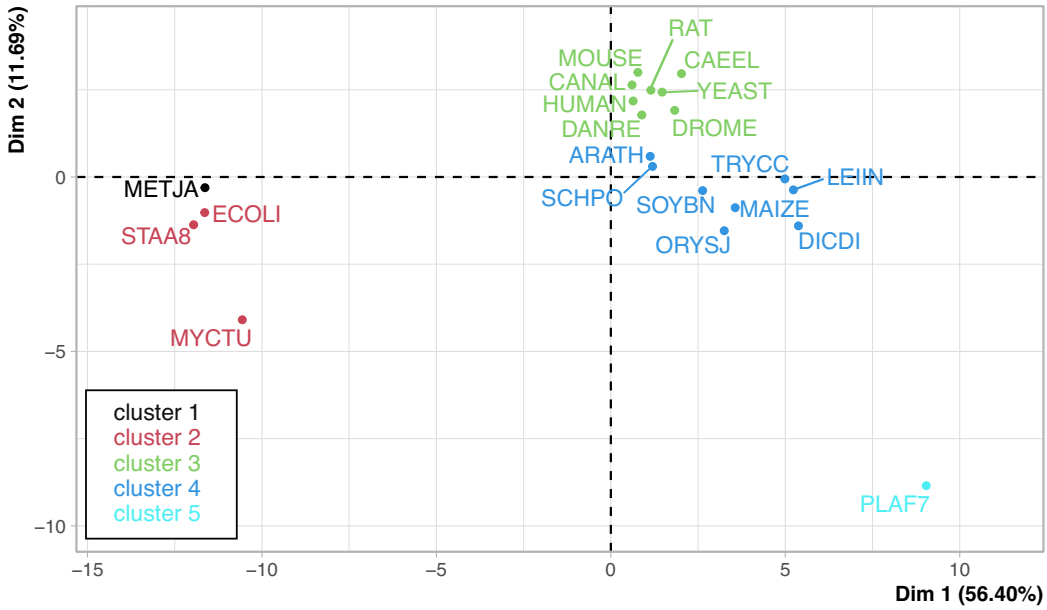

FIGURE 6 Principal component analysis (PCA) for the AFDB v1 21 proteomes according to HCA score and AF2 prediction confidence. The 64 variables correspond to the proportion of residues of foldable segments found in each pLDDT category (very high, confident, low, very low) for each of the HCA score intervals (from −7 to 9, in steps of 1) (as represented in Figure 4). The plot represents the 21 AFDB proteomes in the factor map formed by the first and second principal components (explaining 59.24% of the dataset variability). The correlation circle of the variables is represented in Figure S6. Five clusters of proteomes were defined using hierarchical clustering on the first four principal components (explaining 83% of the dataset variability). ARATH: *Arabidopsis thaliana*, CAEEL: *Caenorhabditis elegans*, CANAL: *Candida albicans*, DANRE: *Danio rerio*, DICDI: *Dictyostelium discoidum*, DROME: *Drosophila melanogaster*, ECOLI: *Escherichia coli*, HUMAN: *Homo sapiens*, LEIIN: *Leishmania infantum*, MAIZE: Zea mays, METJA: *Methanocaldococcus jannaschii*, MOUSE: *Mus musculus*, MYCTU: *Mycobacterium tuberculosis*, ORYSJ: *Oryza sativa*, RAT: *Rattus norvegicus*, SCHPO: *Schizosaccharomyces pombe*, SOYBN: *Glycine max*, STAA8: *Staphylococcus aureus*, TRYCC: *Trypanosoma cruzi*, YEAST: *Saccharomyces cerevisiae*

Although these global trends of relationship between pLDDT and HCA score of foldable segments are conserved across the 21 individual proteomes, several differences should be noticed, as illustrated by five representative patterns in Figure 5B–F. The 21 proteomes have been clustered according to the principal component analysis (PCA) built of their relative proportion of residues of each pLDDT category for foldable segments of each interval of HCA score values (Figure 6). The first principal component explained 56.4% of the total variance and showed that the foldable segments of prokaryotic and eukaryotic proteomes differ for almost all intervals of HCA scores, with higher proportions of very high and very low pLDDT residues, respectively (Figure S6a–d). At second order and according to the second principal component (explaining only 11.7% of the total variance), the proteome of the archaeon *Methanocaldococcus jannaschii* slightly differs from the 3 bacterial proteomes, in particular due to the absence of sequences with HCA scores lower than −4 (corresponding to full disorder) in *M. jannaschii* (Figure 5B,C).

The proteomes of eukaryotes are split into 3 groups, cluster 4 includes the four plant proteomes (*Arabidopsis thaliana*, *Oryza sativa*, *Glycine max*, *Zea mays*), the three *Ascomycota* are separated according to their taxonomy: the two *Saccharomycetaceae* (*Candida albicans* and *Saccharomyces cerevisiae*) belong to cluster 3 whereas *Schizosaccharomyces pombe* is in cluster 4. Cluster 4 is characterized by a higher proportion of low and very low confidence regions in segments with HCA score in]−1;6] (corresponding to globular sequences), and by a lower proportion of very low confidence predictions in segments corresponding to disorder (HCA score lower than −5) (Figure 5D–E). Finally, the proteome of *Plasmodium falciparum* forms another last group, being characterized by predictions of globally lower confidence (Figure 5F). This is consistent with the high level of dark sequences in the Apicomplexa proteomes, for which remote homologs cannot easily been detected for efficient covariation analyses.

In the foldable sequences, we observed two major types of apparent discrepancies between the confidence score of AF2 structural prediction and our classification of structural states based on HCA score. These types indeed deviate from the current assumption that the structure of folded regions would be predicted with high confidence, while disordered regions would not. First, we reported cases of regions of contiguous amino acids with high pLDDT values and low HCA scores, that we suggested to correspond to disorder (785 sequences with length greater than 100 amino acids, corresponding to 20 different proteomes, i.e., at least one sequence in all AFDB v1 proteomes except *Methanocaldococcus jannaschii*; Figure S7a). These regions include the two categories depicted before for nonfoldable segments with high pLDDT values, that is, (a) disordered sequences included in large loops and/or undergoing conditional folding (AF2 then capturing the structure of the complexed state), (b) long, repetitive structure with a particular abundance

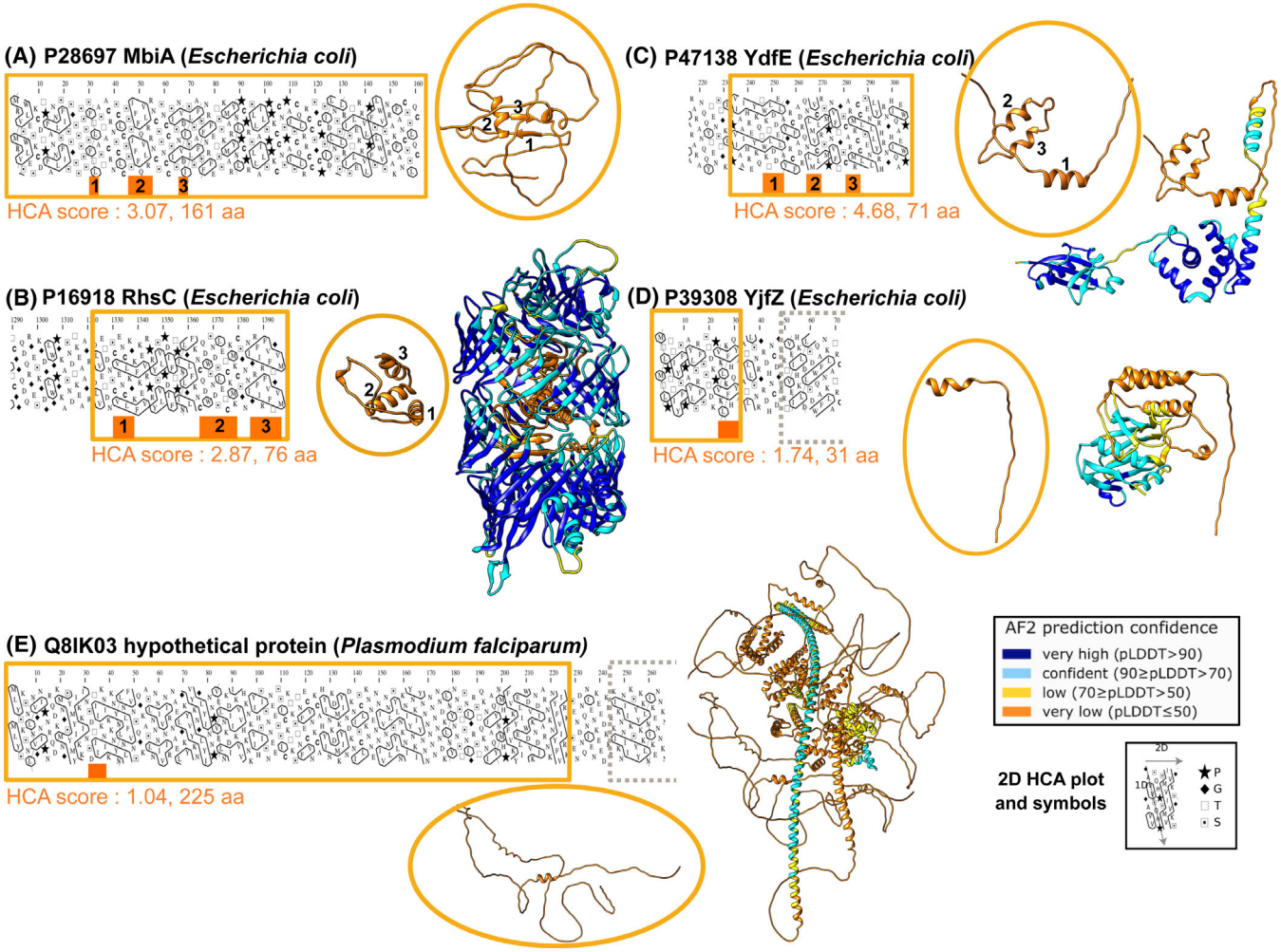

FIGURE 7 Examples of very low AlphaFold2 confidence scores for globular-like domains. The examples shown are segments of contiguous amino acids with very low pLDDT values, fully included in foldable segments with HCA scores typical of well-folded domains (boxed in orange on the HCA plots, with the corresponding HCA scores and segment lengths [in aa] reported below). Other foldable segments of the sequences, taken from UniProt, are boxed in gray. Special symbols used in the HCA representation and the way to read the sequences and RSSs are indicated in the inset. The AF2 3D structure models of the corresponding segments are highlighted with orange circles, extracted from the models of the whole proteins (ribbon representations), colored according to the AF2 per-residue confidence metric. Positions of the RSSs, as predicted by AF2, are reported in color on the HCA plots.

of (alpha-forming) alanine or (beta-forming) serine/threonine and sequences with an ion-dependent folding (Figure S8). Low HCA scores are also observed in some enzymes, which depend on ions for their catalytic activity (e.g., one of the longest regions (224 amino acids) in the human carbonic anhydrase (P00915, HCA score of −0.62)). Apolar but nonstrong hydrophobic amino acids are also particularly abundant in these low HCA score regions within standard globular domains, completing the hydrophobic clusters participating in the hydrophobic core.

Second, we reported cases of regions of contiguous amino acids with very low pLDDT values but with HCA scores typical of folded domains (9 722 sequences with length greater than 100 amino acids, corresponding to 21 different proteomes; Figure S7b,c). Only 54 regions of more than 30 amino acids are found in the *E. coli* proteome (Figure 7), among which a whole protein reported as a recent and rare case of emerging gene by overprinting in this bacterial species (MbiA[74]; Figure 7A), small domains in multidomain proteins (RhsC—Figure 7B, YdfE—Figure 7C) or a protein segment (YjfZ—Figure 7D). RSSs are more or less well predicted by AF2 although their assembly cannot be validated. Of note however is the case of MbiA for which only a part of the RSS is predicted, while the HCA plot clearly indicates several others, associated with hydrophobic clusters typical of beta-strands. This failure of RSS prediction appears recurring in several AF2 predictions of such regions in the *Plasmodium falciparum* proteome, as illustrated in Figure 7E. These results indicate that regions with AF2 very low confidence do not always correspond to disordered regions, in line with studies that have compared AF2 to order/disorder predictors.[72,73]

4 | DISCUSSION

The remarkable progress that has recently been made in the field of structure prediction[75] owes its success to the use of de learning approaches and the consideration of pre-existing knowledge at the protein sequence and structure levels. In particular, the wealth of evolutionary information was leveraged to an optimal level to extract key features of interresidue contacts and distances, which have paved the way to unprecedented levels of accuracy in the prediction.[28,76] Considering evolutionary information was also instrumental in the recent advances made in the field of disorder prediction.[27]

Among the key questions that evolutionary-based methods applied to structure prediction cannot easily address, at least with the expected accuracy, is that of regions lacking known homologous sequences and falling outside family annotations, which constitute the dark proteome.[77–79] A common idea was that the dark proteome is largely made up of IDPs/IDRs, as these are difficult to be characterized by traditional methods of structural biology[18] and moreover harbor less amino acid conservation of their sequences.[80] Contrary to this idea, it was shown that the dark proteome contains an important part of nondisordered sequences, constituting the "unknown unknown".[77,79] A recent survey of the human dark proteome before and after AF2 development has indicated that only a part of these nondisordered sequences was predicted with good accuracy,[81] thereby supporting the ever-present need to develop tools for better characterizing the foldability potential and structural features from the only information of single amino acid sequences and applicable to any proteome, even the darkest ones. It should be noted here that the foldability is linked to the ability of building blocks (i.e., the regular secondary structures) to interact. However, the foldability score does not provide information on how these RSSs interact, which is typically addressed by analysis of residue co-variation, when homologs are available.

The (non)foldability of proteins is encoded in their amino acid sequences, with two global physicochemical patterns, the absolute mean charge and the mean hydropathy, primarily accounting for the differences between two classes.[82] This issue of the foldability potential was addressed here in two steps, based on the consideration of this basic hydrophobic/nonhydrophobic dichotomy, enriched by the information on local structure through the use of a two-dimensional representation of amino acid sequences. Capitalizing on the proven ability of HCA to highlight structural invariants in a context of high evolutionary divergence,[82] we do not explicitly use an hydrophobicity scale especially in order to take into account at the best this dichotomy and the driving role of strong hydrophobic amino acids in the formation of regular secondary structures, regardless of their specific physico-chemical features. The first step of our procedure relies on a binary definition of foldability, with the delineation of homogeneous regions in terms of general properties related to order (foldable segments) or disorder (nonfoldable segments). The second step, focusing on the foldable segments, estimates their degree of foldability in a continuum. This combined approach goes thus beyond a binary and per-residue order/disorder dichotomy and is independent of the consideration of a set of homologous sequences. *pyHCA* thus provides a useful information about the position and global characteristics of structurally homogenous segments, that is, the foldable segments, which contain building blocks that are likely to interact and thus goes beyond the local interactions which can be predicted by considering only isolated hydrophobic clusters. *pyHCA* is in line with the spirit of polymer scaling behavior, which combines hydropathy and charge patterning and is used for characterizing the structural properties of IDPs/IDRs.[83] It is also to be compared to ODiNPred, a sequence order/disorder predictor which uses a deep neural network trained on a database of an experimental, continuous-valued quantification of local disorder based on NMR chemical shifts and considering a large number of sequence features,[33] among which foldable domains as predicted by *SEG-HCA*. The foldable segments defined by *SEG-HCA* are expected to fold spontaneously or conditionally into stable 3D structures through the participation in an hydrophobic core, while nonfoldable segments correspond to full disorder, with the exception of regions whose stably fold without the need of a consistent hydrophobic core, but, for example, depending on ion binding.[49] Conditional foldable segments, having transient residual structures or fold dependent on interactions or environment,[8,11] can generally be distinguished from the autonomous folding units as these segments are often predicted as disordered by current disorder predictors.[42] This category of transient disorder includes short linear motifs (SLIMs)[84] and Molecular Recognition Features (MoRFs),[85] which are generally embedded in large disordered regions. Their intermediate behavior can be highlighted using tools such as *ANCHOR*[86] and visualized with *FELLS*, an estimator of latent structures integrating *SEG-HCA* and *IUPred2* predictions.[87] It is interesting to note that these sequences, which are predicted as disordered but foldable, are globally well predicted by AF2, capturing the folded state, however without detecting their structural plasticity.[73]

In a global way, the HCA scoring scheme introduced here allows to appreciate the degree of foldability of protein segments, reflecting the relative abundance of loop/coil regions (*disorder*) and regular secondary structures (*order*). Thereby, we are able to disentangle the great diversity present within the IDPs/IDRs, which is reflected by a wide range of HCA score values, whereas folded domains are characterized by narrower ranges of values (Figure 1). The distinct behaviors between these two groups have also been evidenced in a recent study using a Gini index, which allows to estimate distribution uniformity.[88]

Based on the HCA score, some foldable segments from DisProt clearly deviate from folded-like segments (intervals a–a′ in Figure 1) while having the capacity to conditionally fold. This is the case, for instance, of the segments shown in Figure 3A–C, with low HCA scores. These segments can thus be easily distinguished from those that are closer to a soluble, globular domain behavior (interval b in Figure 1). IDPs/IDRs found in this last interval of HCA scores are difficult to distinguish from well-folded globular domains based on the only consideration of the HCA score and other features must be considered to refine the analysis. However, one can note that such IDPs/IDRs are generally shorter than typical well-folded globular domains extracted from the SCOPe database, which may explain that they are

unable to fold stably in absence of partners. Indeed, only 48.4% of sequences from this category in DisProt have length greater than 30 amino acids (mean length 81.6 aa), against 85.1% in SCOPe (mean length 133.3 aa). In cases of longer IDPs/IDRs from this category, amino acid composition may help to distinguish them from well-folded globular domains. Indeed, we observed that segments from DisProt of this category (interval b in Figure 1) are enriched in polar amino acids that have been previously described as disorder-promoting (Gln, Lys, Ser, Glu)[89] (Figure S9a). Remarkably, they have composition in strong hydrophobic amino acids comparable to that found in soluble domains. This composition, consistent with high HCA scores, thus defines a specific class of long disordered sequences, reflecting their propensity to fold upon constraint (Figure S10). The example of AF4-AF9 complex, discussed in[49] also suggested that some sequences of IDPs/IDRs might be stabilized in absence of interacting partners by intra-molecular interactions mediated by sequences located at long-range distance in the protein. In this interval b, one can also observe that the composition of well-folded soluble domains is different from that of also well-folded transmembrane beta-barrels, characterized by similar HCA scores values (Figure 1). These have indeed distinctive features such as dyad-repeat patterns and a high abundance in aromatic amino acids at the bilayer interface,[90] allowing their accurate predictions by dedicated tools.[91,92]. Here, we also evidenced an enrichment of beta-barrel foldable segments relative to soluble domain ones in Tyr and Trp, as well as in small and polar amino acids (Gly, Asn, Ser, Thr) (Figure S9a), consistent with previous observations.[90] In the interval c, the DisProt foldable domains can also be distinguished of well-folded alpha-helical membrane domains by their amino acid composition, as the former ones are also enriched in polar/charged amino acids (Asp, Glu, Asn, Gln, Arg, Lys, His, Ser, Thr) (Figure S9b). Examples extracted from this category of DisProt segments highlighted sequences with long hydrophobic clusters (length similar to transmembrane helices), but forming elongated, soluble coiled-coils. First attempts to develop tools for sorting sequences in the intervs b-c and distinguishing between stable structures and conditional order or molten globules, based on these amino acid composition differences, are encouraging. However, further investigations are needed to characterize the building blocks (hydrophobic clusters) of IDPs and sequences linking them, in order to understand the molecular basis of their particular structural behavior, especially in terms of both fuzziness (typified by the co-existence of several minima of free-energy content)[14,93] and frustration.[94,95]

The combination of HCA score and AF2 pLDDT applied to proteome-wide analysis shed light on the relative part of each given proteome in which order is still hidden, corresponding to very low AF2 pLDDT values but with HCA scores typical of globular-like regions (intervals b–c in Figure 1). Examination of particular cases of hidden order indicates that AF2 is able in some situations to predict RSSs, which correlate with hydrophobic clusters (as observed in the *E. coli* sequences presented in Figure 7), nonetheless without confidence in the way they are associated. The overestimation of disorder by AF2 is minimum in the case of prokaryotic proteomes (Figure 8A), where large amount of known 3D structures and sequences are available. In contrast, the accuracy of disorder prediction by AF2 is much lower in case of *Dictyostelium discoideum* and parasitic organisms such as *Trypanosoma cruzi*, *Leishmania infantum* and *Plasmodium falciparum*. For this latter proteome, a percentage of amino acids as high as 40.8% identified in the globular-like category by *pyHCA* correspond to AF2 very low pLDDT scores. Some hypothetical proteins from *Plasmodium falciparum* escape RSS prediction by AF2 and are represented as fully disordered, although they possess hydrophobic clusters with HCA scores typical of well-folded domains. In fact, the higher proportion of amino acids with very low pLDDT values in the *Plasmodium falciparum* proteome (46.0%) cannot be explained in a straightforward way by a higher proportion of disorder, but instead by the compositional bias and low complexity regions leading to mask the order characteristics and to leave a large number of sequences in the dark.[96,97] *pyHCA* is not affected by these biases and estimates a similar foldability trend for *Plasmodium falciparum* as for other eukaryotes such as human (80%–85% aa within foldable segments, Table S1). It allows therefore to unravel the characteristics of the hidden order, as already applied to the identification of hidden actors of the transcription machinery.[98]

Importantly, the reliability of the "order" prediction by pyHCA can unambiguously be demonstrated in cases where the AF2 prediction fails. Indeed, in a joint study,[99] we have focused on long (>30 amino acids) foldable segments with HCA scores typical of soluble, globular-like domains and that, in the corresponding AF2 models, include only amino acids very low pLDDT values and lack regular secondary structures and folded hydrophobic core (random coils). We showed that some of these sequences are reported in the DisProt database as forming transient regular secondary structures (as determined by various biophysical approaches).[99] Moreover, a specific search of these segments against the PDB sequences that were published after the deposit of the AF2 models highlighted a case of an AF2 structureless coil prediction, which was further investigated by NMR and ESR, combined with Rosetta computations.[100] In this study, the observed 3D structure of the intracellular domain (ICD) of human alpha7 nicotinic acetylcholine receptor, which shows conformational plasticity, has proven to be well organized in its resting state (Figure S11). In particular, a loop is anchored onto the intracellular MA helix (merging with the transmembrane helix TM4) via an alpha-helix (h3), which provides with significant hydrophobic contributions to the packing of the MA bundle. The foldable segment contains three hydrophobic clusters, one of which corresponding to the observed helix h3. Thus, these examples of conditional order illustrate the relevance of *pyHCA* for identifying foldability potential in the AF2 structureless coil predictions. This opens the way to reveal novel 3D structures, including some corresponding to unconditional order, which cannot be predicted by AF2, due to lack of homologs and few local patterns corresponding to existing structures.

pyHCA biases mainly rely on regions which stably fold without the need of a consistent hydrophobic core (Figure S5). The analysis of AF2 pLDDT scores outside of foldable segments provides therefore a useful way to evaluate the overestimation of full disorder by *pyHCA* (Figure 8B). At the proteome scale, this bias is minimum for

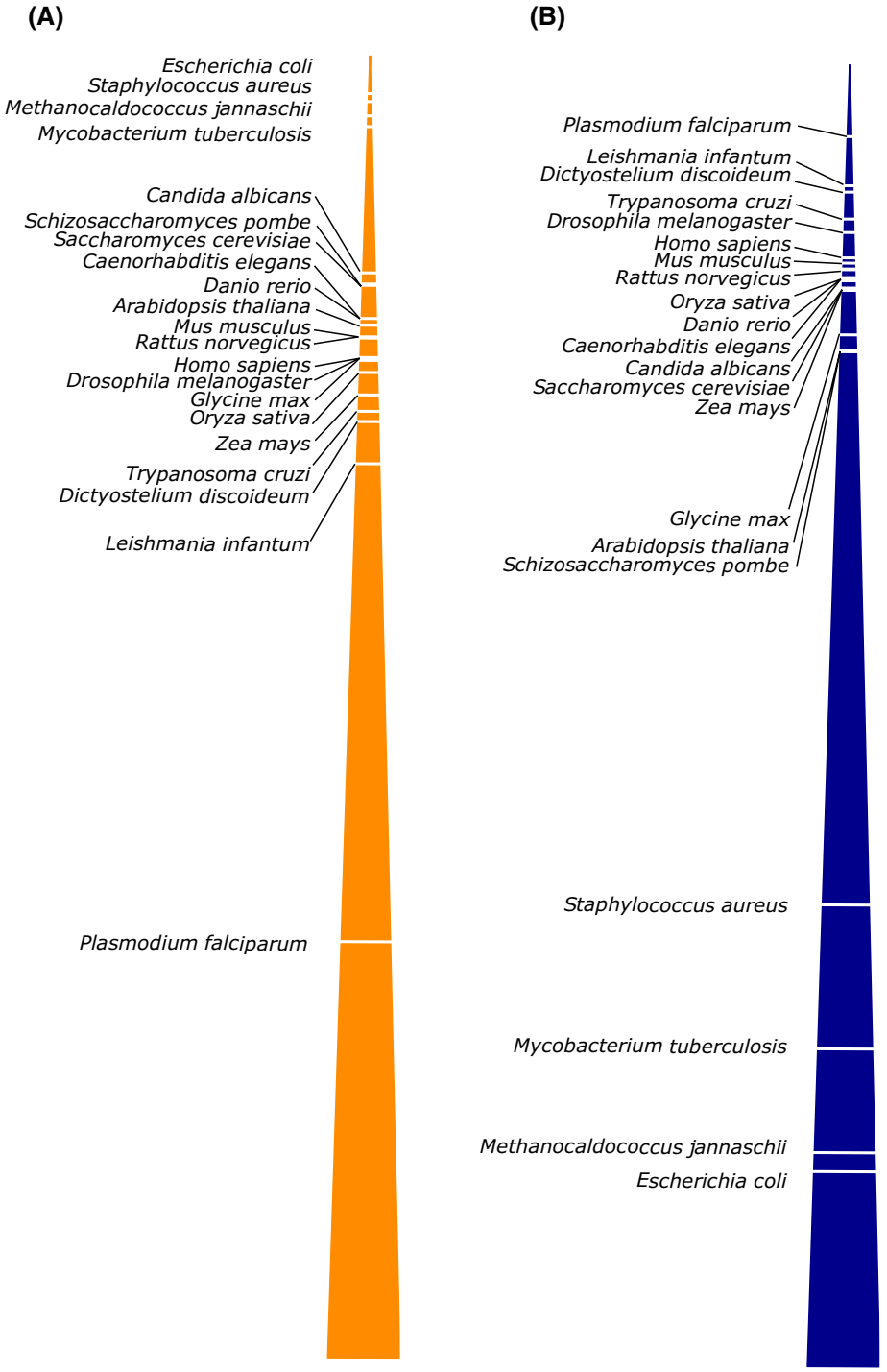

FIGURE 8 Classification of the AFDB v1 21 proteomes according to probable overestimation of disorder by AF2 and pyHCA in foldable and nonfoldable segments, respectively. (A) Proteomes are sorted in a top-down scale (ranging from 0% to 60%) according to the increasing proportion of residues predicted with a very low confidence by AF2 in segments with HCA scores typical of soluble globular HCA scores (category *b* defined in Figure 1). (B) Proteomes are sorted in a top-down scale (ranging from 0% to 60%) according to the increasing proportion of residues found in nonfoldable segments that are predicted with a very high confidence by AF2. For details on the proportion of residues in nonfoldable segments, see Table S1 and Figure S3. The proteome of the archaeon *Methanocaldococcus jannaschii* harbors the lowest number of long regions (length > 30 aa) composed only by residues predicted with a very high confidence by AF2, included in nonfoldable segments. The UniProt sequence accession numbers and the boundaries of these 8 regions are as follows: Q60356 (171–207), Q57673 (15–83), Q58130 (243–276), Q58560 (108–139), Q58991 (253–283), Q60317 (32–64), Q57676 (24–54) and Q58814 (28–61). All have at least one homologous sequence with known 3D structure (data not shown)

Dictyostelium discoideum and the three parasitic organisms (in case of *Plasmodium falciparum*, 3.3% aa outside of foldable segments correspond to a very high pLDDT in AF2 predictions) while ranging to a maximum for the four prokaryotic organisms (51% in *E. coli*). Reminding that these latter proteomes are mostly covered by regions identified as foldable by *pyHCA* (84%–96% aa), this overestimation of disorder by *pyHCA* is however quite low when considering the number of long regions (length higher than 30 aa) capable to fold (e.g., corresponding only to 8 different proteins for *Methanocaldococcus jannaschii*).

Overall, combining *pyHCA* with AF2 provides a revised estimation of the full disorder content in proteomes, corresponding not only to nonfoldable segments which are not well-predicted by AF2, but also to foldable segments with HCA score lower than −4.7. The conditionally folded regions, corresponding to foldable segments with higher HCA score values, are thus discarded from this estimation. Compared

to previous studies,[101] we thus suggest a lower disorder content, as long IDRs (length higher than 30 aa) were found in 12% to 40.4% of long proteins (length higher than 60 aa) in the eukaryotic proteomes in AFDB, and in 0.6%–4.7% for the prokaryotic ones. Our approach also allows to explore the order/disorder content of proteomes in relation to organism ecological traits, in a complementary way to previous works.[24,102] In addition to a specific behavior of parasitic organisms within eukaryotes (see above and Figure 6) that remains to be explained, our study also detected a particularly low disorder content in the proteome of the hyperthermophilic archaeon *Methanocaldococcus jannaschii*, likely related to the high thermal stability constrained by the environmental conditions.

Finally, the scoring scheme offered by *pyHCA* can be used for understanding the protein evolutionary trajectories at the proteome level. It is particularly well-suited to study the structural properties of proteins encoded by de novo emerging genes and how such properties have influenced their early emergence and long-term retention. In particular, it can bring new light to the debate of whether de novo proteins have much intrinsic disorder[103] or are aggregation-prone[104] and whether retention of de novo gene precursor is driven by such properties or remains a stochastic process.[105,106] Several works have already used the concept of foldable segments to investigate such properties,[107,108,109] and a recent work considering a preliminary version of HCA scoring systems has evidenced that most yeast intergenic ORFs contain the elementary building blocks of protein structures.[50] The example of *E. coli* MbiA, an overlapping (protein-coding) orphan gene which has recently evolved by overprinting and was shown to share the structural properties of globular-like domains although not predicted by AF2 (Figure 7A), well illustrates the interest of our approach to decipher structural features in absence of homologs and uncover dark sides of protein evolution.

FUNDING INFORMATION

This work has been supported by the Agence Nationale de la Recherche (grant number ANR-14-CE10-0021, ANR-17-CE12-0016 and ANR-19-CE01-0005) and the Institut National du Cancer (grant number PLBIO14-299). AB was supported by the PhD program of Doctoral School "Complexité du Vivant" (ED515, Sorbonne Université). Analyses were processed with the support of the computer cluster "Plateforme Calcul Intensif Algorithmique" (UMS2700-PCIA) of the Muséum National d'Histoire Naturelle (MNHN). The authors thank Jean-Paul Mornon for critical reading of the manuscript.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY STATEMENT

All data and scripts used in the present work are available in a GitHub repository at https://github.com/DarkVador-HCA/Order-Disorder-continuum. The package pyHCA is provided at https://github.com/DarkVador-HCA/pyHCA under the CeCILL-C license agreement.

ORCID

Isabelle Callebaut https://orcid.org/0000-0003-3124-887X
Elodie Duprat https://orcid.org/0000-0003-4150-6967

REFERENCES

1. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the universe of protein folds. *Annu Rev Biophys*. 2013;42:559-582.
2. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. *Proc Natl Acad Sci U S A*. 2014;111:11691-11696.
3. Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. *PLoS Comput Biol*. 2019;15:e1006969.
4. Schaeffer RD, Kinch LN, Pei J, Medvedev KE, Grishin NV. Completeness and consistency in structural domain classifications. *ACS Omega*. 2021;6:15698-15707.
5. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*. 2014;83:553-584.
6. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114:6589-6631.
7. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2015;16:18-22.
8. Uversky VN. Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J*. 2015;287:1182-1189.
9. Uversky VN. Intrinsically disordered proteins and their "mysterious" (meta)physics. *Front Phys*. 2019;7:10.
10. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 2005;6(3):197-208.
11. Jakob U, Kriwacki R, Uversky VN. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem Rev*. 2014;114:6779-6805.
12. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*. 2017;18:285-298.
13. Wu H, Fuxreiter M. The structure and dynamics of higher-order assemblies: amyloids, signalosomes and granules. *Cell*. 2016;165:1055-1066.
14. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem Sci*. 2008;33:2-8.
15. Lyle N, Das RK, Pappu RV. A quantitative measure for protein conformational heterogeneity. *J Chem Phys*. 2013;139:121907.
16. Wright PE, Dyson HJ. Linking folding and binding. *Curr Opin Struct Biol*. 2009;19:31-38.
17. Borgia A, Borgia MB, Bugge K, et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*. 2018;555:61-66.
18. Bhowmick A, Brookes DH, Yost SR, et al. Finding our way in the dark proteome. *J Am Chem Soc*. 2016;138:9730-9742.
19. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord Proteins*. 2016;4:e1259708.
20. Hatos A, Hajdu-Soltész B, Monzon AM, et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res*. 2020;48:D269-D276.
21. Dosztányi Z. Prediction of protein disorder based on IUPred. *Protein Sci*. 2018;27:331-340.
22. Orlando G, Raimondi D, Codicè F, Tabaro F, Vranken W. Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J Mol Biol*. 2022;434:167579.
23. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 2004;337:635-644.

24. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn*. 2012;30:137-149.
25. Peng Z, Yan J, Fan X, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci*. 2015;72:137-151.
26. Oates ME, Romero P, Ishida T, et al. D2P2: database of disordered protein predictions. *Nucl Acids Res*. 2012;41(D1):D508-D516.
27. Necci M, Piovesan D, Predictors C, Curators D, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. *Nat Methods*. 2021;18:472-481.
28. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.
29. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596:590-596.
30. Akdel M, Pires DEV, Porta Pardo E, et al. A structural biology community assessment of AlphaFold 2 applications. *bioRxiv*. 2021;2021.2009.2026.461876. https://doi.org/10.1101/2021.09.26.461876
31. Wilson CJ, Choy WY, Karttunen M. AlphaFold2: a role for disordered protein/region prediction? *Int J Mol Sci*. 2022;23(9):4591.
32. Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433:167208.
33. Dass R, Mulder FAA, Nielsen JT. ODiNPred: comprehensive prediction of protein order and disorder. *Sci Rep*. 2020;10:14780.
34. Zhang T, Faraggi E, Li Z, Zhou Y. Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem Biophys*. 2013;67(3):1193-1205.
35. Callebaut I, Labesse G, Durand P, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci*. 1997;53:621-645.
36. Eudes R, Le Tuan K, Delettré J, Mornon J-P, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct Biol*. 2007;7:2.
37. Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett*. 1987;224(1):149-155.
38. Lamiable A, Bitard-Feildel T, Rebehmed J, et al. A topology-based investigation of protein interaction sites using hydrophobic cluster analysis. *Biochimie*. 2019;167:68-80.
39. Woodcock S, Mornon J-P, Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng*. 1992;5:629-635.
40. Rebehmed J, Quintus F, Mornon JP, Callebaut I. The respective roles of polar/nonpolar binary patterns and amino acid composition in protein regular secondary structures explored exhaustively using hydrophobic cluster analysis. *Proteins*. 2016;84(5):624-638.
41. Bitard-Feildel T, Lamiable A, Mornon J-P, Callebaut I. Order in disorder as observed by the "hydrophobic cluster analysis" of protein sequences. *Proteomics*. 2018;18:e1800054.
42. Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput Biol*. 2013;9(10):e1003280.
43. Toth-Petroczy A, Palmedo P, Ingraham J, et al. Structured states of disordered proteins from genomic sequences. *Cell*. 2016;167:158-170.
44. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003;31:3701-3708.
45. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*. 2005;347:827-839.
46. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 2005;21:3433-3434.
47. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50:D439-D444.
48. Faure G, Callebaut I. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics*. 2013;29:1726-1733.
49. Bitard-Feildel T, Callebaut I. HCAtk and pyHCA: a toolkit and python API for the hydrophobic cluster analysis of protein sequences. *bioRxiv*. 2018;249995. https://doi.org/10.1101/249995
50. Papadopoulos C, Callebaut I, Gelly JC, et al. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res*. 2021;31:2303-2315.
51. Piovesan D, Necci M, Escobedo N, et al. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res*. 2021;49:D361-D367.
52. Steinegger M, Söding J. MMseqs2: sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35:1026-1028.
53. Fox NK, Brenner SE, Chandonia J-M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*. 2013;42(D1):D304-D309.
54. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*. 2012;40:D370-D376.
55. Shimizu K, Cao W, Saad G, Shoji M, Terada T. Comparative analysis of membrane protein structure databases. *Biochim Biophys Acta Biomembr*. 2018;1860:1077-1091.
56. Ellgaard L, Bettendorff P, Braun D, et al. NMR structures of 36 and 73-residue fragments of the calreticulin P-domain. *J Mol Biol*. 2002;322:773-784.
57. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25:1605-1612.
58. *R: A Language and Environment for Statistical Computing [computer program]*. R Core Team; 2021.
59. Denning DP, Patel SS, Uversky V, Fink AL, Rexach M. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc Natl Acad Sci U S A*. 2003;100(5):2450-2455.
60. Bilokapic S, Schwartz TU. Structural and functional studies of the 252 kDa nucleoporin ELYS reveal distinct roles for its three tethered domains. *Structure (London, England: 1993)*. 2013;21(4):572-580.
61. Huber AH, Weis WI. The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell*. 2001;105(3):391-402.
62. Yang Y, Eichhorn CD, Wang Y, Cascio D, Feigon J. Structural basis of 7SK RNA 5′-γ-phosphate methylation and retention by MePCE. *Nat Chem Biol*. 2019;15(2):132-140.
63. Sá-Moura B, Simões AM, Fraga J, et al. Biochemical and biophysical characterization of recombinant yeast proteasome maturation factor ump1. *Comput Struct Biotechnol J*. 2013;7:e201304006.
64. Uekusa Y, Okawa K, Yagi-Utsumi M, et al. Backbone ^{1}H, ^{13}C and ^{15}N assignments of yeast Ump1, an intrinsically disordered protein that functions as a proteasome assembly chaperone. *Biomol NMR Assign*. 2014;8:383-386.
65. Schnell HM, Walsh RMJ, Rawson S, et al. Structures of chaperone-associated assembly intermediates reveal coordinated mechanisms of proteasome biogenesis. *Nat Struct Mol Biol*. 2021;28:418-425.
66. Kjaergaard M, Teilum K, Poulsen FM. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc Natl Acad Sci U S A*. 2010;107(28):12535-12540.

67. Kjaergaard M, Andersen L, Nielsen LD, Teilum K. A folded excited state of ligand-free nuclear coactivator binding domain (NCBD) underlies plasticity in ligand recognition. *Biochemistry*. 2013;52(10):1686-1693.
68. Xia S, Liu M, Wang C, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res*. 2020;30(4):343-355.
69. Oberer M, Zangger K, Gruber K, Keller W. The solution structure of ParD, the antidote of the ParDE toxin antitoxin module, provides the structural basis for DNA and toxin binding. *Protein Sci*. 2007;16(8):1676-1688.
70. Loris R, Garcia-Pino A. Disorder- and dynamics-based regulatory mechanisms in toxin–antitoxin modules. *Chem Rev*. 2014;114(13):6933-6947.
71. Yan N, Gu L, Kokel D, et al. Structural, biochemical, and functional analyses of CED-9 recognition by the proapoptotic proteins EGL-1 and CED-4. *Mol Cell*. 2004;15(6):999-1006.
72. Aderinwale T, Bharadwaj V, Christoffer C, et al. Real-time structure search and structure classification for AlphaFold protein models. *Commun Biol*. 2022;5(1):316.
73. Alderson TR, Pritišanac I, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv*. 2022; https://doi.org/10.1101/2022.02.18.481080
74. Fellner L, Bechtel N, Witting MA, et al. Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. *FEMS Microbiol Lett*. 2014;350:57-64.
75. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—round XIV. *Proteins Struct Funct Bioinform*. 2021;89(12):1607-1617.
76. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker DI. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. 2020;117:1496-1503.
77. Bitard-Feildel T, Callebaut I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep*. 2017;7:41425.
78. Mistry J, Chuguransky S, Williams L, et al. Pfam: the protein families database in 2021. *Nucl Acids Res*. 2021;49:D412-D419.
79. Perdigão N, Heinrich J, Stolte C, et al. Unexpected features of the dark proteome. *Proc Natl Acad Sci U S A*. 2015;112(52):15898-15903.
80. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res*. 2006;5(4):879-887.
81. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol*. 2022;18:e1009818.
82. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*. 2000;41(3):415-427.
83. Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. *J Phys Chem Lett*. 2020;11(9):3408-3415.
84. Weatheritt RJ, Luck K, Petsalaki E, Davey NE, Gibson TJ. The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics*. 2012;28:976-982.
85. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). *J Mol Biol*. 2006;362:1043-1059.
86. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*. 2009;25:2745-2746.
87. Piovesan D, Walsh I, Minervini G, Tosatto SCEE. FELLS: fast estimator of latent local structure. *Bioinformatics*. 2017;33:1889-1891.
88. Carugo O. Hydrophobicity diversity in globular and nonglobular proteins measured with the Gini index. *Protein Eng des Sel*. 2017;30(12):781-784.
89. Uversky VN, Dunker AK. Understanding protein non-folding. *Biochim Biophys Acta*. 2010;1804(6):1231-1264.
90. Wimley WC. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol*. 2003;13:404-411.
91. Hayat S, Peters C, Shu N, Tsirigos KD, Elofsson A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. *Bioinformatics*. 2016;32:1571-1573.
92. Tian W, Lin M, Tang K, Liang J, Naveed H. High-resolution structure prediction of β-barrel membrane proteins. *Proc Natl Acad Sci U S A*. 2018;115:1511-1516.
93. Miskei M, Horvath A, Vendruscolo M, Fuxreiter M. Sequence-based prediction of fuzzy protein interactions. *J Mol Biol*. 2020;432(7):2289-2303.
94. Freiberger MI, Wolynes PG, Ferreiro DU, Fuxreiter M. Frustration in fuzzy protein complexes leads to interaction versatility. *J Phys Chem B*. 2021;125(10):2513-2520.
95. Malagrinò F, Diop A, Pagano L, Nardella C, Toto A, Gianni S. Unveiling induced folding of intrinsically disordered proteins—protein engineering, frustration and emerging themes. *Curr Opin Struct Biol*. 2022;72:153-160.
96. Pizzi E, Frontalli C. Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Res*. 2001;11:218-229.
97. Hamilton WL, Claessens A, Otto TD, et al. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Res*. 2017;45:1889-1901.
98. Callebaut I, Prat K, Meurice E, Mornon J-P, Tomavo S. Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. *BMC Genomics*. 2005;6:100.
99. Bruley A, Mornon J-P, Duprat E, Callebaut I. Digging into the 3D structure predictions of AlphaFold2 with low confidence: disorder and beyond. *Biomolecules*. 2022;12:1467.
100. Bondarenko V, Wells MM, Chen Q, et al. Structures of highly flexible intracellular domain of human α7 nicotinic acetylcholine receptor. *Nat Comm*. 2022;13(1):793.
101. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*. 2000;11:161-171.
102. Tang Q-Y, Ren W, Wang J, Kaneko K. The statistical trends of protein evolution: a lesson from AlphaFold database. *Mol Biol Evol*. 2022;39:msac197.
103. Basile W, Salvatore M, Bassot C, Elofsson A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol*. 2019;15(7):e1007186.
104. Vakirlis N, Acar O, Hsu B, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun*. 2020;11(1):781.
105. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol*. 2018;2(10):1626-1632.
106. Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic gain and loss of novel transcribed open Reading frames in the human lineage. *Genome Biol Evol*. 2020;12(11):2183-2195.
107. Grandchamp A, Berk K, Dohmen E, Bornberg-Bauer E. New genomic signals underlying the emergence of human proto-genes. *Genes*. 2022;13(2):284.
108. Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J*. 2018;285(14):2605-2625.

109. Watson AK, Lopez P, Bapteste E. Hundreds of out-of-frame remodeled gene families in the *Escherichia coli* pangenome. *Mol Biol Evol*. 2021;39(1):msab329.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bruley A, Bitard-Feildel T, Callebaut I, Duprat E. A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum. *Proteins*. 2022;1-19. doi:10.1002/prot.26441