bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2

T. Reid Alderson[1,2§], Iva Pritišanac[3,4,5§], Alan M. Moses[3], Julie D. Forman-Kay[1,4*]

1. Department of Biochemistry, University of Toronto, ON, Canada M5S 1A8
2. Department of Molecular Genetics, University of Toronto, ON, Canada M5S 1A8
3. Department of Cell and Systems Biology, University of Toronto, Toronto, Canada M5S 35G
4. Program in Molecular Medicine, The Hospital for Sick Children, Toronto, ON, Canada M5G 1X8
5. Gottfried Schatz Research Center for Cell Signaling, Metabolism and Aging, Molecular Biology and Biochemistry, Medical University of Graz, 8010 Graz, Austria

[§] Equal contribution

* Correspondence to forman@sickkids.ca

Keywords: AlphaFold2, AlphaFold Protein Structure Database, intrinsically disordered proteins, intrinsically disordered regions, NMR spectroscopy, conditional folding

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Deep learning-based approaches to protein structure prediction, such as AlphaFold2 and RoseTTAFold, can now define many protein structures with atomic-level accuracy. The AlphaFold Protein Structure Database (AFDB) contains a predicted structure for nearly every protein in the human proteome, including proteins that have intrinsically disordered regions (IDRs), which do not adopt a stable structure and rapidly interconvert between conformations. Although it is generally assumed that IDRs have very low AlphaFold2 confidence scores that reflect low-confidence structural predictions, we show here that AlphaFold2 assigns confident structures to nearly 15% of human IDRs. The amino-acid sequences of IDRs with high-confidence structures do not show significant similarity to the Protein Data Bank; instead, these IDR sequences exhibit a higher degree of positional amino-acid sequence conservation and are more enriched in charged and hydrophobic residues than IDRs with low-confidence structures. We compared the AlphaFold2 predictions to experimental NMR data for a subset of IDRs known to fold under specific conditions, finding that AlphaFold2 tends to capture the folded state structure. We note, however, that these AlphaFold2 predictions cannot detect functionally relevant structural plasticity within IDRs and cannot offer an ensemble representation of IDRs. Nevertheless, AlphaFold2 assigns high-confidence scores to about 60% of a set of 350 IDRs that have been reported to conditionally fold, suggesting that AlphaFold2 has learned to identify conditionally folded IDRs, which is unexpected, since IDRs were minimally represented in the training data. Leveraging this ability to discover IDRs that conditionally fold, we find that up to 80% of IDRs in archaea and bacteria are predicted to conditionally fold, but less than 20% of eukaryotic IDRs. Our results suggest that a large majority of IDRs in the proteomes of human and other eukaryotes would be expected to function in the absence of conditional folding.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The accurate prediction of protein structures from amino-acid sequences has been a long-term goal in biology (Anfinsen 1973; Baker & Sali 2001). If truly accurate for all targets, protein structure prediction would circumvent the laborious structure determination process, accelerate the pace of biological discovery, and enable novel applications in biotechnology and medicine via the rational design of structures with specific functions (Kuhlman & Bradley 2019). The biennial Critical Assessment of Structure Prediction (CASP) competition (Moult et al. 1995) has stimulated many developments in the field of protein structure prediction, including the successful implementation of co-evolutionary restraints derived from multiple sequence alignments (MSAs) and machine learning protocols in CASP12 (Moult et al. 2018; Schaarschmidt et al. 2018). CASP14 brought a revolutionary advancement: the AlphaFold2 team at DeepMind produced more models with atomic-level accuracy than ever before in the history of CASP (AlQuraishi 2021; Jumper et al. 2021a,b). The second-best scoring prediction software in CASP14 led to the RoseTTAFold structure prediction platform, which was released in open-source format and contained a webserver for ease of access (Baek et al. 2021). Subsequently, DeepMind predicted the structures for 98.5% of proteins in the human proteome (Tunyasuvunakool et al. 2021). The predicted structures were made publicly available, in collaboration with the European Bioinformatics Institute, via the AlphaFold Protein Structure Database (AFDB) (https://alphafold.ebi.ac.uk/) (Varadi et al. 2021).

An unexpected effect of the AFDB is that it visually demonstrates the prevalence of intrinsically disordered regions (IDRs) in the human proteome. IDRs are predicted to comprise *ca.* 30% of the human proteome; play important cellular roles in transcription, translation, and signaling (Dyson & Wright 2005); and are enriched in proteins associated with neurological and other diseases (Uversky et al. 2008). Moreover, it has recently become evident that IDRs contribute to and modulate the formation of many *in vivo* biomolecular condensates via effects on liquid-liquid phase separation (Borcherds et al. 2021; Martin & Holehouse 2020). Numerous disease-associated mutations are found in IDRs (Vacic et al. 2012), including mutations implicated in autism-spectrum disorder (ASD) and cancer (Tsang et al. 2020), and aberrant phase separation involving IDRs has been linked to diseases such as ALS, ASD, and cancer (Alberti & Dormann 2019; Tsang et al. 2020), highlighting the need to understand the structural and biophysical impact of these mutations. At the structural level, IDRs are defined by a lack of stable secondary

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and tertiary structures and rapid interconversion between different conformations (Van Der Lee et al. 2014; Wright & Dyson 2015). Because of their rapid dynamics, IDRs are not amenable to high-resolution structure determination methods and are frequently removed or not observed in structures determined by X-ray crystallography and cryo-electron microscopy. By contrast, AlphaFold2-generated structural models contain the entire protein sequence, including IDRs (Ruff & Pappu 2021), and one can now visualize predictions for the significant fraction of the proteome that was previously "dark" and unobservable (Bhowmick et al. 2016).

IDRs, however, do not adopt the static structures that are depicted in the AFDB (Ruff & Pappu 2021). Instead, IDRs populate an ensemble of interconverting conformations that depends strongly on the primary structure (Das & Pappu 2013), and the properties of these ensembles directly impact on the functions of IDRs (Borcherds et al. 2014; Conicella et al. 2020; Das et al. 2016; Iešmantavičius et al. 2014; Kim et al. 2021; Maltsev et al. 2012; Milles et al. 2015; Mittag et al. 2008; Sugase et al. 2007; Zosel et al. 2018). However, experimentally determined structural information for IDR conformational ensembles constitutes only a tiny fraction of the available data for structured proteins (Lazar et al. 2021; Varadi et al. 2014), and such ensembles are not deposited in the Protein Data Bank (PDB), an online repository that contains more than 150,000 high-resolution structures of biomolecules (Burley & Berman 2021), data which were mined to create AlphaFold2 (Jumper et al. 2021a) and RoseTTAFold (Baek et al. 2021). Protein structure-prediction programs that make use of available data in the PDB, therefore, will be biased by the relatively few available structures of IDRs, which typically involve those IDRs that fold upon binding to an interaction partner (Smith et al. 2021; Wright & Dyson 2009). The presence of IDR structures in the PDB skews the view of other functional states of IDRs and provides no information for myriad other IDRs that do not fit the "folding-upon-binding" paradigm (Borgia et al. 2018; Fuxreiter 2019; Murthy & Fawzi 2020).

Nuclear magnetic resonance (NMR) spectroscopy is well suited to an ensemble-based structural characterization of IDRs at atomic resolution (Jensen et al. 2014; Konrat 2014; Mittag & Forman-Kay 2007). The intrinsic dynamics of IDRs often lead to long-lived NMR signals that can be exploited to collect high-quality data (Malki et al. 2021; Sugase et al. 2007; Theillet et al. 2016). Indeed, a battery of NMR experiments have been applied to probe the conformations of IDRs and residual structure therein (Bertoncini et al. 2005; Dyson & Wright 2021; Eliezer 2007; Kakeshpour et al. 2021; Mantsyzov et al. 2014;

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Salmon et al. 2010), with dedicated software programs focused on integrating NMR and other biophysical methods to determine ensemble representations of IDPs that best agree with the experimental data (Bottaro et al. 2020; Choy & Forman-Kay 2001; Gomes et al. 2020; Krzeminski et al. 2013; Lincoff et al. 2020; Ozenne et al. 2012; Salmon et al. 2010). However, both the integrative structural biology approach used to determine ensemble representations of IDRs and the NMR-driven determination of residual structure or secondary structure propensity in IDRs are often accessible only to specialists. Such data are not deposited in the PDB, which is used to train and validate deep-learning models, such as AlphaFold2. Finally, because AlphaFold2 was trained on a subset of the PDB that excluded NMR structures (Jumper et al. 2021a), NMR data offer a unique validation metric to assess the accuracy of predicted AlphaFold2 structures in solution, as recently demonstrated (Robertson et al. 2021; Zweckstetter 2021).

Due to the extent and functional importance of IDRs and their under-representation in databases used in the training of AlphaFold2 and subsequent models (Baek et al. 2021; Evans et al. 2021; Jumper et al. 2021a), it is essential to assess such predictions in the context of available structural data and our current understanding of the structural propensities and functional mechanisms of IDRs. In this work, we took a first step in quantifying the extent of predicted structures in the human AFDB within regions that are known or predicted to be IDRs.

Here, we show that thousands of IDRs are predicted by AlphaFold2 to be folded with high ($70 \leq x < 90$) or very high (≥ 90) predicted local difference distance test (pLDDT) scores (Mariani et al. 2013), which measure the confidence in the predicted structures (Jumper et al. 2021b). We find that, compared to IDRs with low pLDDT scores, the amino-acid sequences of IDRs with high pLDDT scores are enriched in charged and hydrophobic residues, show more positional conservation, and have more alignment matches to sequences in the PDB. However, only 4% of high-scoring IDR sequences have alignment matches in the PDB, indicating that structural templating is not the reason that AlphaFold2 confidently folds these IDRs. For a subset of IDRs that fold under specific conditions and have been extensively characterized by NMR spectroscopy, we find that the AlphaFold2 structures of these IDRs resemble the conformation of the folded state. Moreover, we observed that AlphaFold2 assigns confident pLDDT scores to nearly 60% of the residues derived from *ca*. 350 IDRs that are known to conditionally fold, suggesting that AlphaFold2 can discover disordered regions that fold upon binding or modification. Therefore, we propose that IDRs with

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

high pLDDT scores may fold in the presence of specific binding partners or following post-translational modifications, which we refer to as conditional folding. However, we note that the AFDB does not capture the multiplicity of conformational states accessible to IDRs or their dependence on context (interactions or modification), and hence fails to inform on the plasticity that is crucial to the function of IDRs. We suggest that the ensemble representation of IDRs is key area for future research in computational protein structure predictions. Finally, we compare predictions of conditional folding in eukaryotes, bacteria, and archaea and find that the latter kingdoms show much higher proportions of conditionally folding IDRs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

In this work, we focus on the structural predictions that are available in the AFDB, which contains pre-computed AlphaFold2 models that can be easily visualized and downloaded for offline inspection. The AFDB democratizes access to otherwise computationally expensive calculations that require both sufficient hardware and computational literacy to perform. Moreover, the lightweight versions of AlphaFold2 that have been implemented as Jupyter Notebooks on Google Colaboratory (Mirdita et al. 2021; Tunyasuvunakool et al. 2021) do not use the same MSAs that were used in the AFDB and thus do not produce the same result. For IDRs, we find that the structural predictions from these lightweight versions are generally of lower quality and do not agree well with AFDB (**Supplementary Figure 1**). As such, we focus henceforth exclusively on structural predictions within the AFDB.

We first analyzed the distribution of per-residue pLDDT scores in the human AFDB (**Figure 1A**). The histogram of pLDDT scores shows a clear bimodal distribution, with the local maxima of each distribution centered around values of 100 and 35 (**Figure 1A**). The majority of residues, accounting for 62.6% of the proteome, have pLDDT scores greater than 70 (**Figure 1A**), which is defined to be the lower threshold for a “confident” score. The remaining 37.4% of residues in the proteome have pLDDT scores below 70 (“low”), while 27.8% of residues have scores below 50 (“very low”). Thus, while a significant percentage of residues have “confident” or “very confident” pLDDT scores, suggesting that the predicted structures of regions are expected to be accurate, there also exists a sizeable fraction of residues that have “low” to “very low” pLDDT scores (**Figure 1A**), indicative of low-accuracy structures that should not be interpreted quantitatively.

Predicted human IDRs with high pLDDT scores in the AFDB

To determine how many IDRs in the AFDB have high pLDDT scores that reflect high confidence structural predictions, we extracted the predicted IDRs from the human AFDB (**Figure 1B**). We used the state-of-the-art sequence-based predictor of intrinsic disorder, SPOT-Disorder (Hanson et al. 2017), to calculate the predicted disorder propensities for each protein in the human proteome (**Figure 1B**). A total of *ca.* 3.5 million residues are predicted to be disordered, totalling 32.8% of the human proteome, consistent with literature values (Necci et al. 2021) (**Supplementary Table 1**). We then investigated the proteins that

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comprise the lower end of the distribution of pLDDT scores within the AFDB. Naively, one might assume that the *ca.* 32.8% of predicted residues in IDRs would be embedded in the *ca.* 37.4% of residues with pLDDT scores below 70, because IDRs should have "low" or "very low" confidence structural predictions. Moreover, an inverse correlation between pLDDT scores and predicted disorder was previously noted (Tunyasuvunakool et al. 2021), with pLDDT scores even reported to perform well as a new predictor of intrinsic disorder (Necci et al. 2021; Tunyasuvunakool et al. 2021).

However, when we isolated the pLDDT scores from residues that localize to SPOT-Disorder predicted IDRs (**Figure 1B**), we found that IDRs also have a bimodal distribution of pLDDT scores (**Figure 1A**). Of the *ca.* 3.5 million predicted disordered residues, *ca.* 14.3% (*i.e.*, ca. 500,000 residues in total) have "confident" pLDDT scores greater than or equal to 70 (**Figure 1A**, **Figure 1C**, **Supplementary Table 2**). When the pLDDT threshold is increased to greater than or equal to 90 ("very confident"), there are more than 160,000 residues that remain, accounting for 4.5% of the total number of disordered residues (**Figure 1A, Figure 1C**). This analysis indicates that there is a significant fraction of SPOT-Disorder-predicted IDRs in the human AFDB that have high-confidence structural predictions (**Figure 1A**, **Supplementary Figure 2**), and hence the general assumption that all IDRs have low pLDDT scores is incorrect.

To ensure that the "confident" and "very confident" scores associated with SPOT-Disorder-predicted IDRs are not the result of poor or biased disorder predictions, we extracted the pLDDT scores for IDRs in the DisProt database of experimentally validated IDRs (Quaglia et al. 2021). There is a total of 932 human IDRs in DisProt, yielding over 300,000 residues that can be used as a direct comparison to the SPOT-Disorder predictions. We find that the distribution of pLDDT scores for IDRs in DisProt shows an even higher proportion of residues with "confident" scores greater than or equal to 70 than the SPOT-Disorder-predicted IDRs: more than 29.5% of the DisProt IDRs have confident pLDDT scores (**Supplementary Figure 2**). Thus, the SPOT-Disorder predictions do not contain an artificially inflated fraction of residues with high pLDDT scores. In addition, we checked if the high pLDDT scores originate from a few disordered residues that are immediately adjacent to structured domains. To this end, we filtered the predicted IDRs with confident pLDDT scores (greater than or equal to 70) for those with consecutive regions of disorder. We found that over 50% of IDRs with high pLDDT scores come from stretches of 24 or more consecutive disordered residues, with nearly 10% arising from IDRs that have 100 or more

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consecutive disordered residues (**Supplementary Figure 2**). Finally, we filtered the list of IDRs to extract those that have 10 and 30 or more consecutive residues with "confident" ("very confident") pLDDT scores. We identified 8084 (2452) and 2644 (923) IDRs that respectively match these criteria.

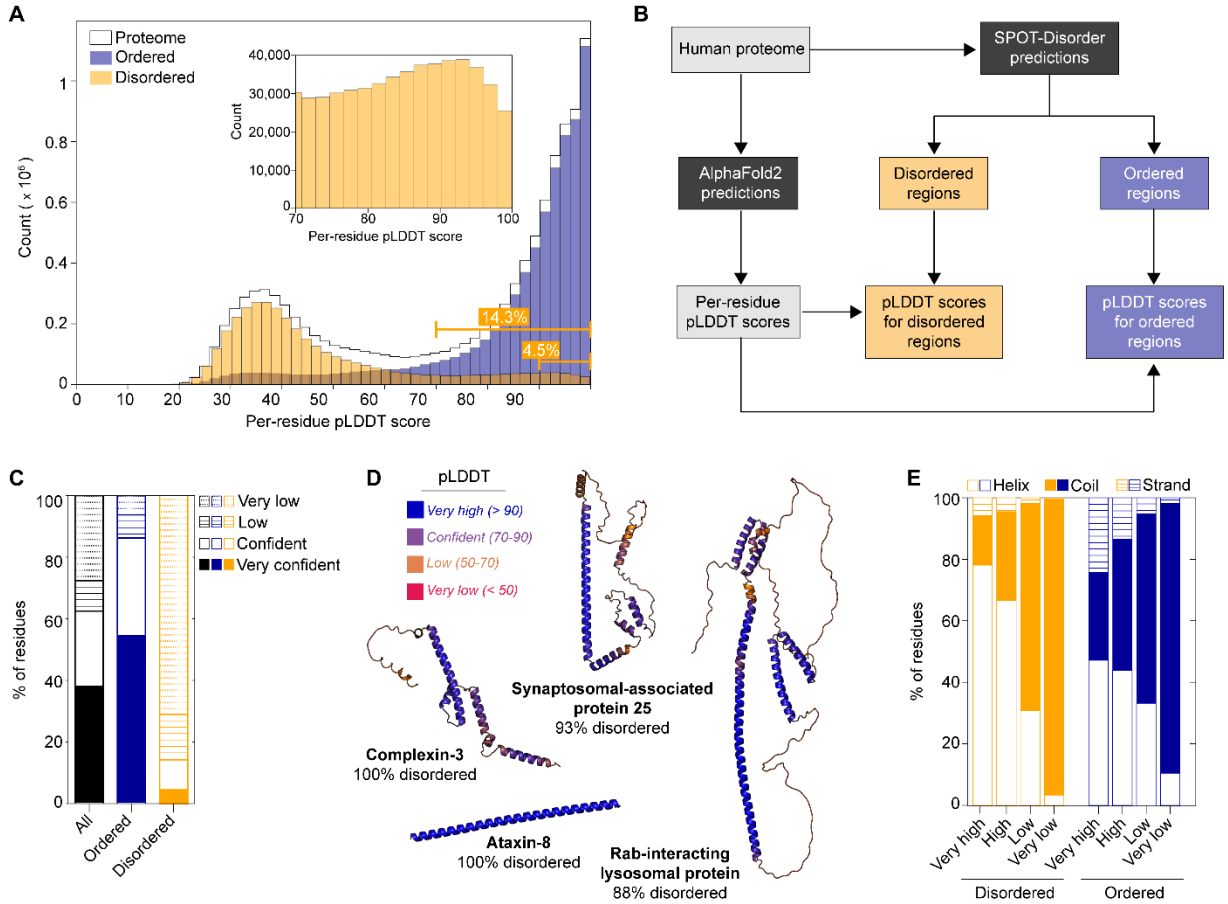

Figure 1. **Predicted IDRs in the human proteome that have confident structures in the AFDB.** (**A**) Histogram of per-residue pLDDT scores in the human proteome (black) compared with the predicted disordered (orange) or ordered (blue) regions. The inset shows an expansion of the predicted disordered regions between pLDDT scores of 70 to 100. The cumulative percentage of predicted disordered residues with scores greater than or equal to 70 and 90 are indicated in the lower right. (**B**) Flowchart outlining the analysis presented in (A). The human proteome was separated into predicted ordered and disordered regions using a sequence-based predictor of disorder (SPOT-Disorder). Per-residue pLDDT scores were obtained for each protein in the AFDB and split into predicted ordered or disordered regions based on the SPOT-Disorder results. (**C**) Stacked bar graph showing the percentage of residues in the human proteome (black) that have very low (< 50; dotted lines), low (50 ≤ x < 70; horizontal lines), confident (≤ 70 x < 90; empty), and very confident (≤ 90; filled) pLDDT scores. The corresponding plots are included for SPOT-Disorder-predicted disordered residues (orange) and ordered residues (blue). (**D**) Example structures in the AFDB significantly overlapping SPOT-Disorder-predicted IDRs, with the percentage of predicted disordered residues of the total listed. The AFDB structures have been color-coded by pLDDT scores as indicated. (**E**) DSSP-determined secondary structure content of the predicted disordered (orange) and ordered (blue) regions, which were grouped as a function of pLDDT threshold, defined above.

From the list of proteins that contain predicted IDRs equal to or longer than 30 residues with high pLDDT scores, we selected a handful of examples for structural analysis in the AFDB (**Figure 1D**). In

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

particular, we identified proteins that are predicted to be predominantly disordered by SPOT-Disorder yet are assigned very high pLDDT scores in the AFDB. For example, the protein ataxin-8 (UniProt ID: Q156A1) contains an initial Met residue followed by 79 Gln residues and is predicted to be fully disordered (**Figure 1D**). However, the AlphaFold2 model indicates that ataxin-8 forms a single α-helical structure with pLDDT scores greater than 90 for every residue in the helix (**Figure 1D**). Similarly, the proteins complexin-3 (UniProt ID: Q8WVH0), synaptosomal-associated protein 25 (SNAP25, UniProt ID: P60880), and Rab-interacting lysosomal protein (RILP, UniProt ID: Q96NA2) all adopt highly α-helical structures with very high pLDDT scores and various degrees of tertiary interactions (**Figure 1D**), despite being predicted by SPOT-Disorder to be predominantly disordered.

Given that the above examples were α-helical structures, we computed the secondary structure content for every model in the AFDB to assess the structural properties of IDRs with high-confidence pLDDT scores. This analysis revealed primarily helical conformations in the "high" and "very high" confidence IDR structures (**Figure 1E**). When compared to ordered regions, the predicted IDRs are significantly enriched in helical conformations at the expense of coils and strands (**Figure 1E**, **Supplementary Table 3**). In addition, we note that the predicted IDRs with low confidence scores still exhibit significant secondary structure content: over 32% of residues with low pLDDT scores are assigned to regions of secondary structure as compared to 38% in ordered regions (**Figure 1E**, **Supplementary Table 3**). In the very low scoring predicted IDRs, the percentage of residues in regions of secondary structure dramatically diminishes to only 3.4% as compared to 12% in ordered regions (**Figure 1E**, **Supplementary Table 3**).

Overall, the analysis of secondary structure content in predicted IDRs with high pLDDT scores shows an enrichment in helical conformations. Moreover, the selected examples in **Figure 1D** all have at least one long, extended α-helix that is not stabilized by tertiary contacts. These so-called single α-helix (SAH) domains are well known in the literature and are estimated to exist in 0.2-1.5% of human proteins (Barnes et al. 2019; Swanson & Sivaramakrishnan 2014), with the formation of SAHs dependent on stabilizing *i* to *i*+4 salt bridges between charged side chains (Marqusee & Baldwin 1987). SAH domains are enriched in E, K, and R residues, which form in various repeats and sum to nearly 80% of the residues in SAH sequences, while the remaining 20% of the sequences derive from A, D, L, M, and Q residues, (Simm & Kollmar 2018). The long α-helix in ataxin-8, which is composed almost entirely of Q, represents another

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

form of an experimentally observed SAH. NMR analyses recently showed that regions containing homo-repeats of Q (polyQ) also form SAHs via side-chain *i* to main-chain *i*-4 hydrogen bonds, although the helical structure decays toward the C-terminal region of the polyQ tract and is dependent on non-Gln hydrogen-bond acceptors at the N-terminus of the helix (Escobedo et al. 2019). Thus, although the sequences of the protein regions in **Figure 1D** would not be classified as *bona fide* SAH domains, the presence of SAH-like structures in the AFDB does not necessarily mean that these structures are non-plausible or non-physical. It is likely that the SAH-like structural regions for predicted IDRs in the AFDB represent structural states that are stabilized upon interactions or modification, including a combination of canonical SAHs and long α-helices that form stabilizing inter-molecular contacts (*e.g.*, coiled coils).

Comparing NMR data and AlphaFold2 structures for experimentally characterized IDRs

Given that our above structural analyses relied on sequence-based predictions of intrinsic disorder, we asked whether the structures of IDRs with high pLDDT scores show correspondence with experimentally determined structural propensities for a subset of IDRs. To this end, we focus on three IDRs/IDPs that have been characterized in detail by NMR spectroscopy (Alderson et al. 2018; Bah et al. 2015; Bodner et al. 2009; Dawson et al. 2020; Demarest et al. 2002; Ebert et al. 2008; Eliezer et al. 2001; Mantsyzov et al. 2015; Marsh et al. 2006; Salmon et al. 2010; Theillet et al. 2016). Two of the model proteins, α-synuclein (UniProt ID: P37840) and 4E-BP2 (UniProt ID: Q13542), are full-length IDPs, whereas the third protein, ACTR or NCoA3 (UniProt ID: Q9Y6Q9), is a small IDR that is part of a larger protein with folded domains and other longer IDRs. The AlphaFold2-generated structures of the three proteins (**Figure 2A**) vary from all helical (α-synuclein, ACTR) to a mixture of strand and helix (4E-BP2). For each structure, the pLDDT scores in the regions of secondary structure range from “high” to “very high” (**Figure 2B**), suggestive of atomic-level accuracy and an overall high level of confidence in the structures (Jumper et al. 2021a; Varadi et al. 2021). Next, we checked the predicted disorder propensity using four different sequence-based predictors of intrinsic disorder, and we found that either two (4E-BP2, ACTR) or three (α-synuclein) of the four programs predicted that these proteins would be predominantly ordered (**Figure 2C**). Thus, without additional experimental evidence, an AFDB user who relies on the overlap between sequence-based

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disorder prediction software and the (confident) AFDB structure would likely assume that the IDR/IDP under investigation folds into the predicted structure.

However, we find that the AlphaFold2 models of these IDRs/IDPs and some of the sequence-based predictors of disorder disagree with the experimental NMR data (**Figure 2D**). It is well known that ^{13}Cα and ^{13}Cβ chemical shifts are sensitive reporters on the secondary structure of a protein (Cornilescu et al. 1999; Spera & Bax 1991; Wishart et al. 1991). For each residue in the protein, the expected chemical shifts for a fully disordered state can be subtracted from the measured chemical shifts. These so-called secondary chemical shifts (corrected for neighboring residues) provide residue-level information regarding the secondary structure of a protein, including the fluctuating, fractional secondary structure of disordered regions, which can be quantified using software programs such as δ2D, SSP, and CheSPI (Camilloni et al. 2012; Marsh et al. 2006; Nielsen & Mulder 2021). If the AFDB structures were correct, the expected secondary chemical shifts would reveal long stretches of structure with SSP values near 1 or -1 for a fully stabilized α-helix or a β-strand, respectively (**Figure 2D**). By contrast, the experimental NMR data for each of three proteins in **Figure 2** show that there is no stable secondary structure and only a fractional preference to populate secondary structure (**Figure 2D**). Thus, a user without knowledge of the disordered nature of these proteins from experiment could erroneously trust the confident AlphaFold2 models (**Figure 2A**, **Figure 2C**) and use the lack of predicted disorder (**Figure 2B**) as a cross-validation method to justify the structures.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

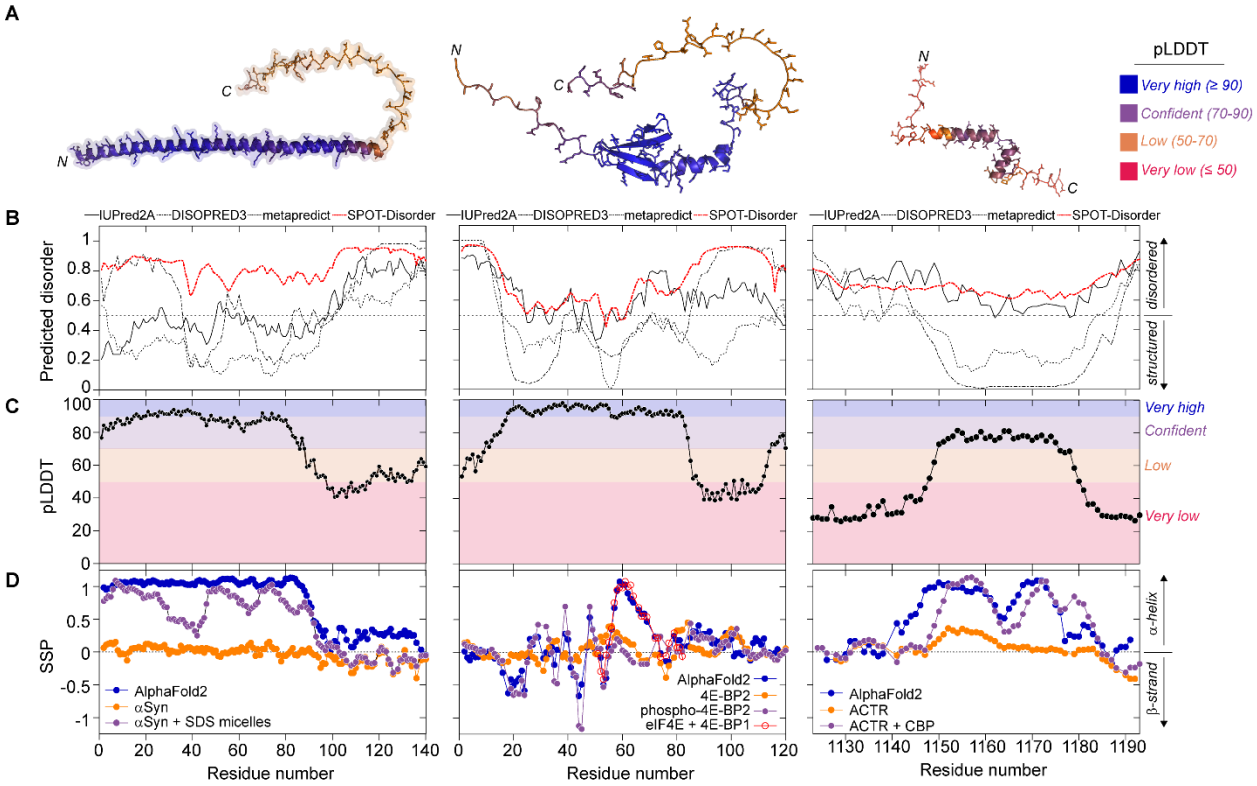

Figure 2. **Examples of three IDRs with high pLDDT scores that conditionally fold and have been characterized by NMR spectroscopy.** (**A**) AlphaFold2-predicted structures of three IDPs/IDRs that have been extensively characterized by NMR spectroscopy. The structures are color coded by pLDDT scores. From left to right: α-synuclein, 4E-BP2, and ACTR. The N- and C-termini of each protein are indicated. (**B**) Sequence-based prediction of disorder for the three IDPs/IDRs. Four different programs were used: IUPred2A, DISOPRED3, metapredict, and SPOT-Disorder. Only SPOT-Disorder correctly predicts the disordered nature of all three IDPs/IDRs. (**C**) Per-residue pLDDT confidence scores derived from the AlphaFold2 structures. The thresholds are defined according to the AFDB: very high (90-100), confident (70-90), low (50-70), and very low (< 50). (**D**) NMR ^{13}Cα and ^{13}Cβ chemical shift-derived secondary structure propensity (SSP) (Marsh et al. 2006). For the AlphaFold2 structures (blue) and the 4E-BP2 peptide bound to eIF4E (PDB ID: 3am7), NMR chemical shifts were back-calculated from the structure using SPARTA+ software (Shen & Bax 2010). The unbound/unmodified IDPs/IDRs (orange) show very little preferential secondary structure (α-synuclein) or modest populations of helix (4E-BP2, ACTR). By contrast, the binding to SDS micelles (α-synuclein), phosphorylation (4E-BP2), binding to eIF4E (4E-BP2), or the binding to CBP (ACTR) induces the formation of stable secondary structure (purple) that is in better agreement with the AlphaFold2 structures.

Interestingly, however, there are correlations between the AFDB structures of these IDRs/IDPs and their experimentally defined conformations under specific conditions. For example, the N-terminal *ca.* 100 residues of α-synuclein fold into a long α-helix in the presence of lipid vesicles (Bodner et al. 2009), and the AFDB structure reflects this lipid-bound conformation (**Figure 2A**). There is no high-resolution structure of the lipid-bound state, so no side-by-side structural comparison can be made; NMR data are only available for α-synuclein bound to SDS micelles (**Figure 2D**, purple). 4E-BP2 folds into a four β-strand structure upon multi-site phosphorylation (PDB ID: 2mx4) (Bah et al. 2015), and the AFDB structure has correctly identified

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the β-strands and the intermolecular contacts: the heavy-atom RMSD is 0.35 Å upon alignment of the β-strands from T19-D55 in the experimental structure to the AFDB model. Confusingly, however, an additional helix in the AFDB model is present in residues R56-R62, followed by a short turn and then a 3_{10}-helix between residues P66-Q69. These additional helices resemble those seen in crystal structures of non-phosphorylated 4E-BP2 and 4E-BP1 bound to eukaryotic translation initiation factor 4E (eIF4E) (PDB IDs: 3am7, 5bxv). For ACTR, a helix-turn-helix motif is present in the AFDB structure, whereas a three-helix structure is formed upon binding to CBP (PDB: 1kbh). ACTR provides a particularly useful test case because the experimental structure (PDB ID: 1kbh) was determined by NMR spectroscopy, and AlphaFold2 was not trained on NMR structures (Jumper et al. 2021a).

Finally, we note that fractional secondary structure in the unbound form of an IDR does not necessarily correlate with the secondary structure in the AFDB model of the IDR. For example, in the unbound form of ACTR, there is an α-helix populated to approximately 40% between residues 1150-1163 (**Figure 2D**, orange), which closely matches the position of the first α-helix in the AFDB (**Figure 2D**, blue) and experimental structures (**Figure 2D**, purple). However, the second and third α-helices in the experimental structure are not appreciably formed in the unbound state or the AFDB model (**Figure 2D**, orange). The lack of clear correlation between fractional secondary structure in an isolated IDR and stabilized structure in complexes has been noted previously (Marsh et al. 2010).

These comparisons show that the high-confidence AFDB structures of IDRs do not reflect the conformational ensemble sampled by the unbound or unmodified form of the IDR. Instead, the AlphaFold2 structures appear to resemble the conditionally folded state of the IDR. Moreover, in the case of 4E-BP2, the AlphaFold2 model has combined structural features from two different conditionally folded forms of the protein that do not coexist: one structure forms upon multi-site phosphorylation (β-strand-rich) and the other upon binding to eIF4E (helical). In this case, the AlphaFold2 structure of 4E-BP2 obscures the molecular mechanism of the protein (see section below).

AlphaFold2 structures of experimentally characterized IDRs resemble the conditionally folded state but do not capture structural plasticity

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our analysis of IDRs/IDPs with extensive NMR data showed that AFDB models with high pLDDT scores might reflect a conformation of the IDR/IDP that only forms under specific conditions. We thus examined the AlphaFold2 structures for an additional four IDRs/IDPs that are known to fold upon binding to interacting partners and have high-resolution structures of the complex in the PDB. The structures for two of these complexes were determined by X-ray crystallography (p27: 1jsu, SNAP25: 1kil) and two were determined by NMR spectroscopy (HIF-1α: 1l8c, CITED2: 1p4q). A comparison of the experimental structures (**Figure 3A-D**) with those in the AFDB (**Figure 3E-H**) shows an overall high structural similarity (**Figure 3I-L**). For the two examples with very confident AFDB structures (p27, SNAP25), the heavy atom root-mean-squared-deviations (RMSDs) are 0.5 and 2.1 Å (**Figure 3E,F,I,J**). Even for some structures that have a mixture of very confident and low pLDDT scores (CITED2, **Figure 3F, 3G**), or only low pLDDT scores (HIF-1α, **Figure 3H**), the overall architecture of the AFDB structure resembles that of the experimental structure, with RMSD values of 1.6 (**Figure 3K**) and 5.0 Å (**Figure 3L**), respectively. Taken together, these analyses suggest that the AFDB structures formed by IDRs with high pLDDDT scores are likely capturing some structural features that form in the presence of specific interactions. In the case of an IDR with very low pLDDT scores (HIF-1α), the presence of secondary structure appears to correlate with the final bound-state conformation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

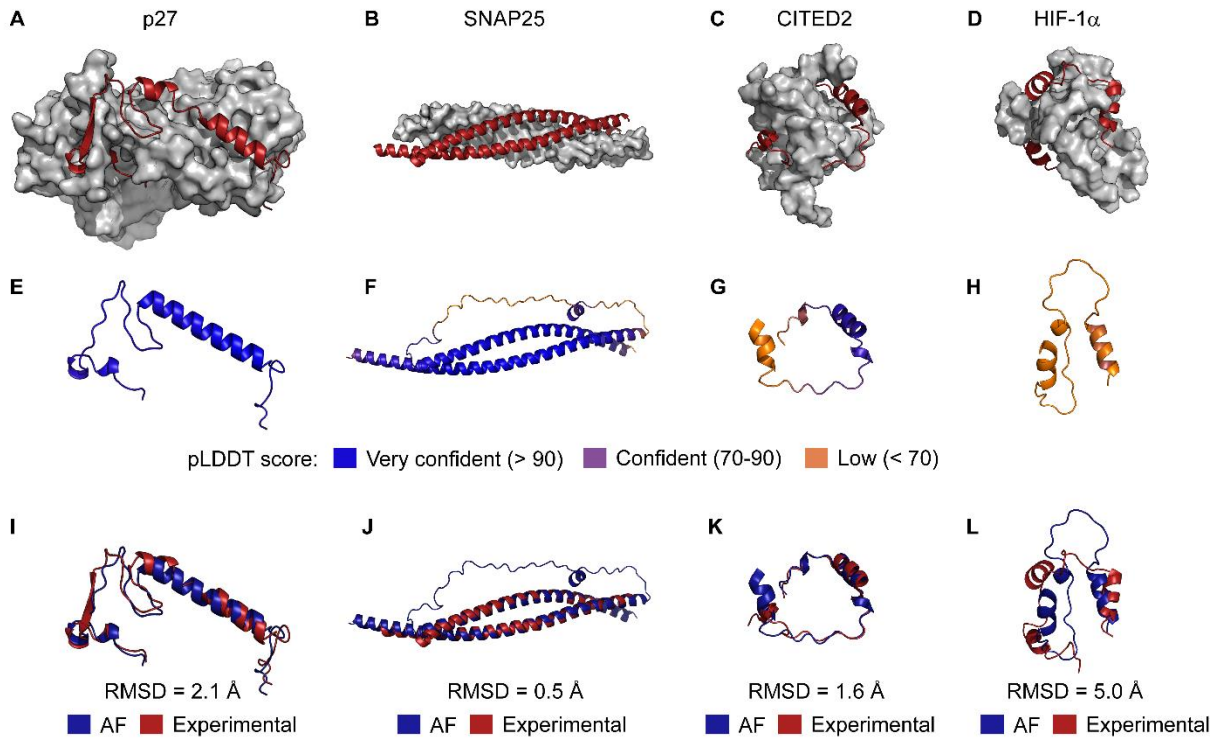

Figure 3. Structures of IDRs in the AFDB correlate with experimentally determined structures of the IDRs bound to interaction partners. (**A, B, C, D**) experimental structures for the listed IDRs/IDPs (red) bound to an interacting folded domain (grey surface representation). The PDB ID codes are 1jsu, 1kil, 1l8c, and 1p4q, respectively. (**E, F, G, H**) the predicted structures in the AFDB for the listed IDRs/IDPs in panels A-D. These structures have been color-coded by per-residue pLDDT scores, with blue, purple, and orange respectively corresponding to very confident (≥ 90), confident (70-90), and low (< 70) scores. (**I, J, K, L**) comparison of the experimental structures from panels A-D with the predicted structures in the AFDB from panels E-H. Experimental structures are colored red and AFDB structures blue. The heavy-atom RMSD upon alignment of secondary structure elements is indicated.

If AlphaFold2 structures of IDRs reflect the conditionally folded state, we were interested to determine if these structures can inform on the molecular mechanisms of IDRs. The interconversion between different structural forms is essential for IDR function, and it is well known that IDRs can bind to multiple interaction partners via different interfaces or motifs, often times forming unique structural elements in the process. The development of small molecules that target the binding sites for IDRs, as well as the IDRs themselves, is an active area of pharmaceutical research (Metallo 2010). However, given that the AFDB contains only a single structure of a given IDR, it is important to emphasize that these AlphaFold2 models are but one possible conformation of the IDR, especially in the context of an IDR binding to an interaction partner. We show below how the AlphaFold2 models obscures insight into molecular

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mechanisms of IDRs by imposing a single structure onto an IDR/IDP that exists in an equilibrium between many accessible conformations.

To demonstrate this, we selected three IDRs with multiple experimentally determined structures in which the IDR has folded into a different conformation. Some of the experimental structures of these IDPs or IDRs from the cystic fibrosis transmembrane conductance regulator (CFTR; UniProt ID: P13569), SNAP-25, and 4E-BP2 show good agreement with the AlphaFold2 models (**Figure 4A-C**). For example, the regulatory (R) region of CFTR is a long IDR that is heavily phosphorylated with several regions that adopt residual helical propensity (Baker et al. 2007; Bozoky et al. 2013b). The weak, multivalent inter- and intramolecular interactions between the R region and different binding partners regulate the activity of CFTR in a phosphorylation-dependent manner (Bozoky et al. 2013b). The AlphaFold2 model of the portion of the CFTR R region immediately following the first nucleotide-binding domain (NBD1), called the regulatory extension (RE), shows close agreement with its conformation in a crystal structure of NBD1 and the RE (**Figure 4D**). However, another structure of NBD1 shows the RE interacting with a different interface on NBD1, with the orientation of RE with respect to NBD1 dramatically altered, despite almost no changes to the structure of NBD1 itself (Bozoky et al. 2013a).

Another example is provided by SNAP-25, which is an IDP that folds into a helical bundle in SNARE complexes (**Figure 4B**) that have important functions in membrane fusion during synaptic vesicle exocytosis. The AlphaFold2 model of SNAP-25 correctly identifies the N- and C-terminal soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) motifs that form a four-helix bundle in the ternary SNARE complex involving SNAP-25, synaptobrevin, and syntaxin (**Figure 4E**). However, SNAP-25 is also a substrate for Clostridal neurotoxins (CNTs), which are zinc-dependent endopeptidases that cause the diseases tetanus and botulism by specifically cleaving SNARE proteins and impairing neuronal exocytosis (Schiavo et al. 1992). The crystal structure of the botulinum neurotoxin serotype A (BoNT/A) protease bound to SNAP-25 reveals an extensive interface involving the C-terminal SNARE motif of SNAP-25 (Breidenbach & Brunger 2004) (**Figure 4I**). An α-helix is formed in BoNT/A-bound SNAP25 by residues D147-M167, while the remaining residues G168-G204 of SNAP-25 are bound to BoNT/A in a coil conformation, with a small β-strand (K201-L203) formed as well (Breidenbach & Brunger 2004) (**Figure 4I**). By contrast, in the SNARE complex bound to complexin-1, SNAP25 forms a long α-helix that encompasses

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

residues S140-M202 (Chen et al. 2002) (**Figure 4B**). These three structures of SNAP-25 provide an illustrative example: given the very high confidence in the AlphaFold2 structure of SNAP-25, one could assume that an ordered-to-disordered transition is required for SNAP-25 to bind to BoNT/A in the experimentally observed conformation, and that SNAP-25 assembles into SNARE complexes as a rigid body with minimal structural changes. Both of these assumptions are in stark contrast to the known disordered-to-ordered transition that occurs both upon binding to BoNT/A and formation of the SNARE complex. Thus, the molecular mechanism of SNAP-25 function, and its proteolytic cleavage in disease, are obscured by the high-confidence AlphaFold2 model.

Finally, we examined the case of 4E-BP2, an IDP that is a regulatory binding protein for the eIF4E, with experimental structures of segments of the protein in the 5-site phosphorylated state and in the non-phosphorylated state bound to eIF4E (**Figure 4C, 4F, 4J**), as discussed in the section above. The AlphaFold2 structure of residues A16-P72 of 4E-BP2 contains a β-sheet followed by α- and 3_{10}-helices (**Figure 4F**), whereas the experimental structure in phosphorylated 4E-BP2 contains the β-sheet for residues T19-D55 followed by a coil region (Bah et al. 2015) (**Figure 4C**). The AlphaFold2 model correctly places the β-strands and accurately identifies the orientations of each strand relative to one another (**Figure 4F**). However, the helical secondary structure elements are only observed when non-phosphorylated 4E-BP2 binds to eIF4E (PDB ID: 3am7) (**Figure 4J**). Upon binding to eIF4E, a truncated peptide from 4E-BP2 was shown to fold into an α-helix between residues D55-D61 (**Figure 4J)**. The related protein 4E-BP1, for which more complete structural information is available when bound to eIF4E (PDB ID: 5bxv), forms an α-helix between residues D55-R62 followed by a short turn and a 3_{10}-helix between residues P66-Q69. The phosphorylation-induced folded state inhibits the binding of eIF4E that is otherwise extremely tight for the unmodified protein. The β-strand-rich structure of the AlphaFold2 model of 4E-BP2 is incompatible with the binding to eIF4E, which requires the disordered non-phosphorylated state known to fractionally sample helical structure in the segment that forms a stable α-helix in complex with eIF4E (Lukhele et al. 2013). Thus, the 4E-BP2 AFDB structure reflects a strange mixture of the ordered landscape of the protein, combining that found in the presence of PTMs with that stabilized in the absence of PTMs but in the presence of a protein binding partner, and confounding understanding of the mechanism of phospho-regulation of translation initiation (Bah et al. 2015).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Given that predicted IDRs/IDPs with high or very high pLDDT scores are more widespread than initially thought (**Figure 1A**), it is critical to be able to determine if the AFDB structures are correct. The comparisons outlined above relied on known structural knowledge or assigned chemical shifts that were obtained from detailed analysis of multidimensional NMR spectra, which requires considerable time and effort. We suggest how an integrative biophysical approach can rapidly assess the accuracy of predicted structures for IDRs with high or very high pLDDT scores in the AFDB by comparing experimental measurements with those back-calculated from an input structure (**Supplementary Figure 3**, **Supplemental Appendix**).

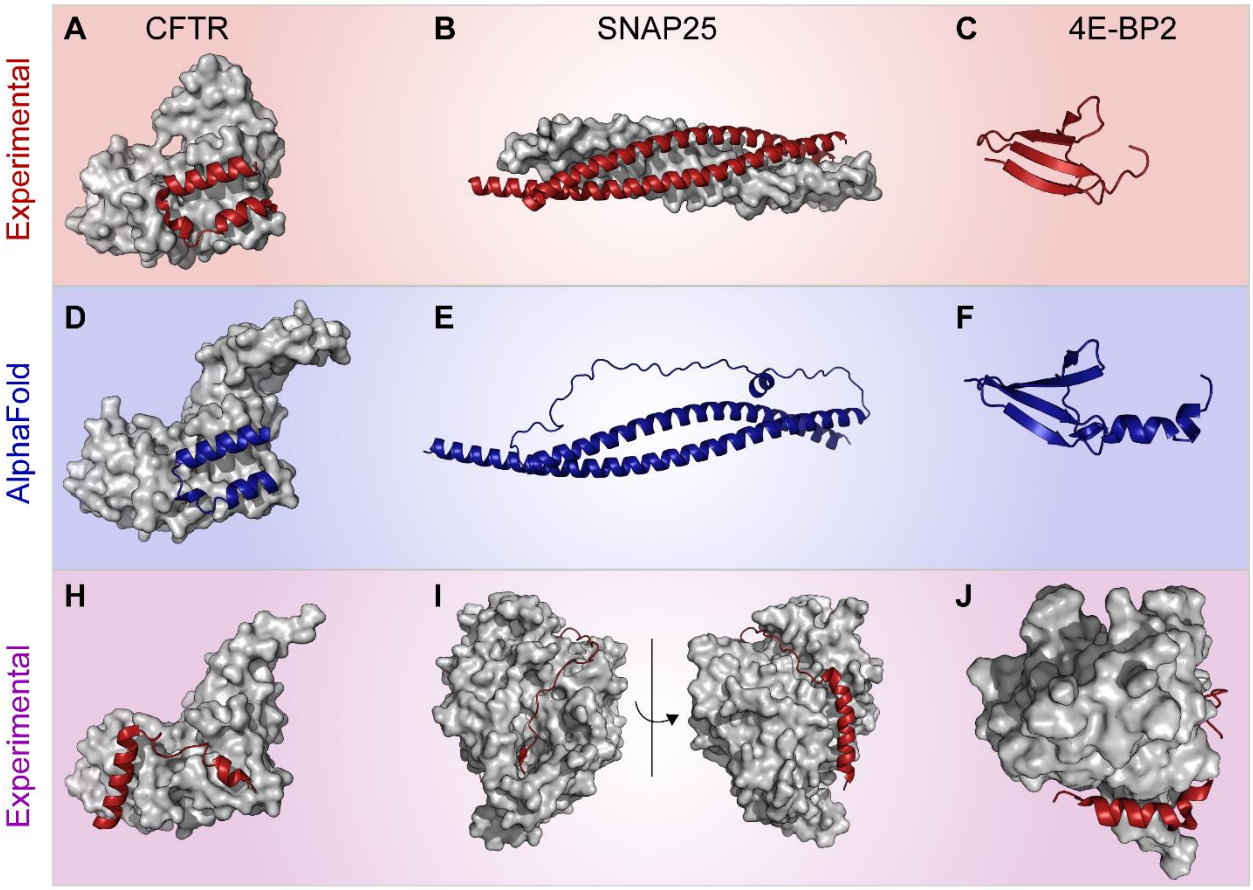

Figure 4. **The AFDB does not capture the inherent structural plasticity of IDRs/IDPs.** Shown here (top) are three examples of IDRs/IDPs that have experimentally determined structures when bound to interacting partners. (**A**) The disordered regulatory extension (RE) of human CFTR (residues L636-S670) bound intramolecularly to NBD1 (residues S388-S635; PDB ID: 1r0x). Note that the N-terminal region of NBD1 contains a cloning artifact and so residues S388-I393 differ from the AlphaFold2 model. (**B**) The N- and C-terminal SNARE motifs of human SNAP25 (residues S10-L81 and G139-W204, respectively) bound to rat VAMP2 (residues S28-N93), rat syntaxin-1A (residues L192-D250), and rat complexin-1 (residues K32-I72; PDB ID: 1kil). (**C**) Phosphorylated 4E-BP2 (residues P18-R62; PDB ID: 2mx4. (**D-F**) The AlphaFold2-predicted structures (blue) of the CFTR RE (residues P638-S670), SNAP25 (M1-G206), and 4E-BP2 (A16-P72) show excellent correspondence with the experimental structures in A-C. (**H-J**) However, the CFTR RE,

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SNAP25, and 4E-BP2 have also been captured in a different conformation as those in panels A-C (red). (**H**) CFTR RE (P638-L671) intramolecularly bound in a different orientation to NBD1 (S388-Q637; PDB ID: 1xmi). (**I**) SNAP25 C-terminal SNARE motif (residues M146-G204) bound to BoNT/A) protease (residues P2-R425; PDB ID: 1xtg). (**J**) A peptide from 4E-BP2 (residues T50-P84) bound to eIF4E (residues H33-K206; PDB ID: 5bxv).

Rigid-body docking with confidently AlphaFold2-predicted IDP/IDR structures

The above examples collectively demonstrate how high-confidence AlphaFold2 structures of IDRs/IDPs, which can offer insight into various structures that are accessible to the IDR/IDP, may also obscure the molecular mechanisms of these disordered regions. Next, we investigated this problem from the other side: if high-confidence structures of IDRs/IDPs are capturing the bound/modified states of IDRs/IDPs, then can such structures be used with protein-protein docking software to obtain structural models of IDR/IDP complexes bound to globular domains? If so, then high-confidence AlphaFold2 structures of IDRs/IDPs could be used with the goal of identifying the interfaces of IDR/IDP-globular domain complexes. As we have shown above (**Figure 2, Figure 3, Figure 4**), for predicted IDRs/IDPs with high pLDDT scores, AlphaFold2 appears to capture the conditionally folded states of such IDPs/IDRs. Therefore, if the AFDB structure of the IDR/IDP is already in a bound-state conformation, it may be beneficial to use the AlphaFold2 model of the IDR/IDP as a starting structure for a protein-protein docking analysis with a known or putative globular domain interactor. Molecular docking with IDRs/IDPs could be valuable for understanding molecular mechanisms, interaction sites, and binding interfaces in the absence of experimentally determined structures of the IDR/IDP-globular domain complex.

However, we illustrate how this approach can be seriously flawed. As an example, we used an experimentally determined complex structure of the CITED2 transactivation domain (TAD) bound to the CBP TAZ1 domain (PDB 1p4q), since this enables a comparison to the experimental structure of the complex. Indeed, as we showed above, the AlphaFold2-predicted structure of the CITED2 TAD closely resembles the CBP-bound form (**Supplementary Figure 4C, 4G, 4K**), with a heavy-atom RMSD between the experimental and AlphaFold2 structures of only 1.6 Å for the entire region and 1.0 Å when aligning the helices only (**Supplementary Figure 4A**). As a control, we first extracted the individual chains for the CITED2 TAD and the CBP TAZ1 in the experimental structure, and rigid-body docked these two chains with protein-protein docking software (**Supplementary Figure 4B**). Reassuringly, the lowest-energy docked structure agreed with the experimental complex (heavy-atom RMSD for residues N216-F259: 0.9

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Å), indicating that this strategy could work for AlphaFold2 structures that have atomic-level accuracy to the conditionally folded state of the IDR.

Next, we rigid-body docked the AlphaFold2 structure of the CITED2 TAD onto the experimentally determined structure of the globular domain (TAZ1) (**Supplementary Figure 4C**). The results from this simple docking exercise are quite striking: even though the AlphaFold2 structure of the CITED2 TAD closely agrees with the experimental structure of CITED2 (1.6-Å RMSD for all residues or 1-Å RMSD for helices-only), the orientation of the C-terminal helix in the AFDB structure is shifted by 90° relative to the experimental structure (**Supplementary Figure 4A**). The rotation of the C-terminal helix in CITED2 causes a steric clash with the CBP TAZ1 domain that dramatically alters the lowest-energy structure of the docked complex (**Supplementary Figure 4C**). Thus, if one naively used the AlphaFold2 structure of the CITED2 TAD to dock this structure into its interacting globular domain, the resultant molecular model would be seriously flawed.

Predicted IDRs with high pLDDT scores are enriched in charged and hydrophobic residues that promote secondary structure

Our above structural analyses focused on specific IDRs/IDPs that conditionally fold and have been experimentally characterized in both the free/disordered and bound/ordered states. We next sought to understand why AlphaFold2 is folding IDRs/IDPs into high-confidence structures. To this end, we extracted all SPOT-Disorder-predicted IDRs in the AFDB for bioinformatic analyses. We hypothesized that the predicted IDRs with high pLDDT scores might manifest for one or a combination of reasons: (1) global amino-acid sequence differences in comparison to the predicted IDRs with low pLDDT scores, (2) relatively high positional sequence conservation (i.e., “high quality” multiple sequence alignments (MSA)), and (3) the enrichment of high-pLDDT IDR sequences in the PDB. The first possibility would reflect a differential “folding propensity” that is inherently encoded in the amino-acid sequences of high vs. low pLDDT-scoring IDRs, whereas the latter two possibilities would influence the AlphaFold2 prediction confidence due to the depth of the MSAs (2) or sequence similarity to the structures from the PDB used in training (3) (Jumper et al. 2021a,b). Given the relatively poor coverage of IDRs in the PDB (Quaglia et al. 2021) and the poor

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

positional alignability for most IDRs (Colak et al. 2013; Nguyen Ba et al. 2012; Zarin et al. 2019, 2021), it is plausible that some combination of all three of the aforementioned possibilities could contribute to high pLDDT scoring IDRs.

To gain insight into these possibilities, we first computed the amino-acid frequencies for each of the following three categories: predicted disordered regions with low pLDDT scores below 50 ($IDR_{low\ pLDDT}$), predicted disordered regions with high pLDDT scores greater than or equal to 70 ($IDR_{high\ pLDDT}$), and predicted ordered regions (ordered). We hypothesized that the amino-acid frequencies in $IDR_{low\ pLDDT}$ should reflect the sequence biases found in disordered regions, *i.e.* an enrichment in some charged (D, E, K), polar (Q, S, T), small (G), and helix-disrupting (P) residues (Quaglia et al. 2021). Indeed, the difference between $IDR_{low\ pLDDT}$ and ordered regions ($\Delta_{ordered}$) shows that $IDR_{low\ pLDDT}$ sequences are enriched in the expected residues that are known to promote disorder (**Figure 5A**, **Supplementary Figure 5**).

We next compared $IDR_{low\ pLDDT}$ and $IDR_{high\ pLDDT}$ regions (Δ_{IDR}) to determine if there are global differences in the sequences of these IDRs that are encoded within per-residue pLDDT scores. Surprisingly, we found that $IDR_{high\ pLDDT}$ sequences are significantly enriched in E, K, Q, and R residues relative to $IDR_{low\ pLDDT}$ sequences (**Figure 5A**). This is evident by the positive-to-negative sign change (E, Q, R) or a large negative value (K) when comparing $\Delta_{ordered}$ vs. Δ_{IDR}. Furthermore, $IDR_{high\ pLDDT}$ sequences have relatively fewer of some canonical disorder-promoting residues (*e.g.*, P, S, T, D, G) and more order-promoting residues (*e.g.*, C, F, I, L, V, W, Y). Nonetheless, $IDR_{high\ pLDDT}$ sequences still resemble IDR sequences when analyzed by mean net charge and mean hydropathy (**Supplementary Figure 6**), which is a sequence metric that identifies IDRs from ordered regions. However, the $IDR_{high\ pLDDT}$ sequences show a much broader distribution in both the mean hydropathy and mean net charge dimensions than the $IDR_{low\ pLDDT}$ sequences (**Supplementary Figure 6**). Thus, although the $IDR_{high\ pLDDT}$ sequences contain more disorder-promoting residues than ordered regions (**Figure 5B**), $IDR_{high\ pLDDT}$ sequences appear to have a mixture of both order- and disorder-promoting residues.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

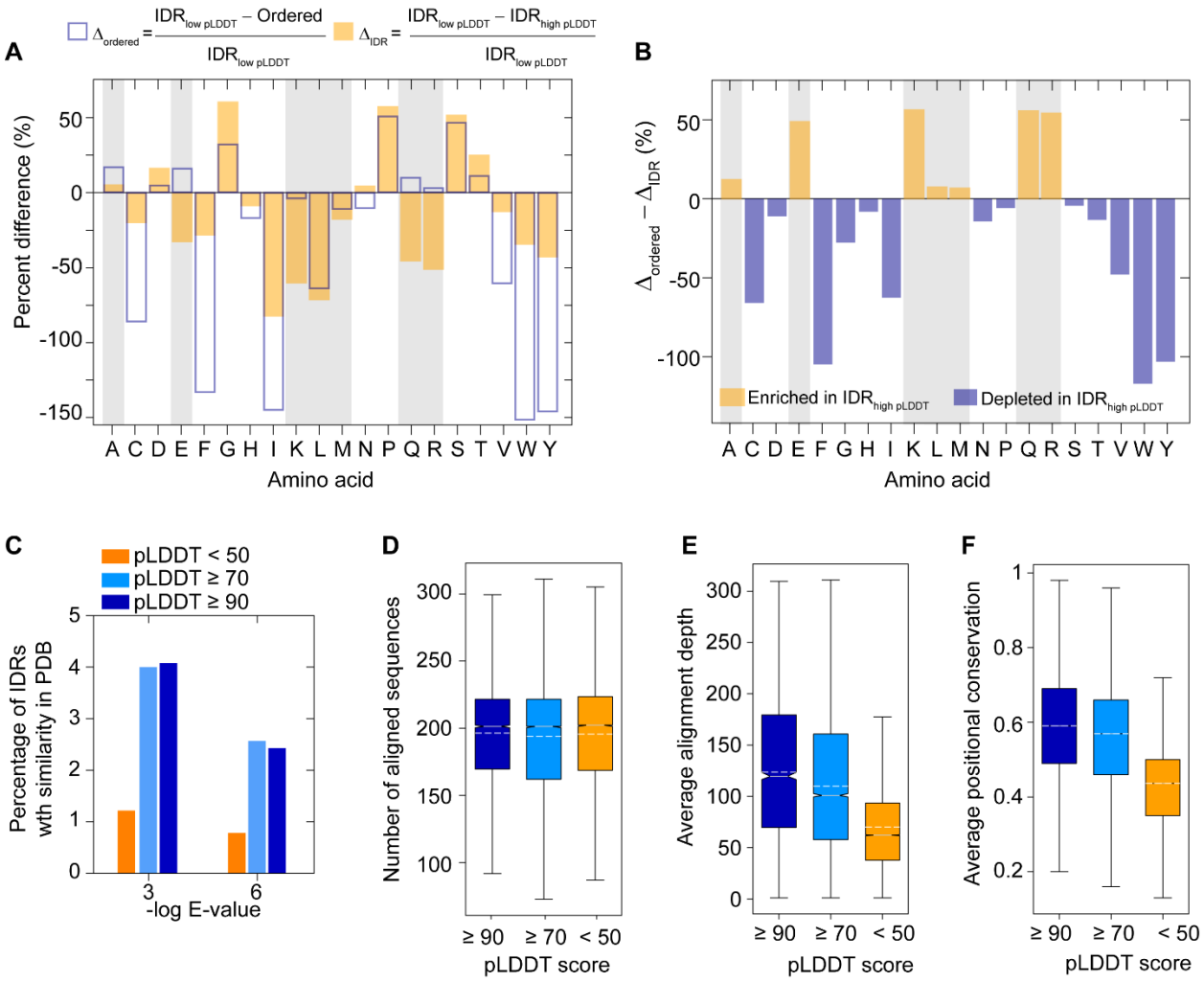

Figure 5. **Bioinformatics analysis of predicted IDRs in the AFDB with high pLDDT scores.** (**A**) Amino-acid percentages in the regions of predicted order and disorder, with the disordered regions further separated into those with confident pLDDT scores greater than or equal to 70 ($IDR_{high\ pLDDT}$) and those below 50 ($IDR_{low\ pLDDT}$). Shown here is the percent change in the relative amino-acid percentages (eq 1) for $IDR_{low\ pLDDT}$ and either ordered regions ($\Delta_{IDR\text{-}order}$, empty blue bars) or $IDR_{high\ pLDDT}$ ($\Delta_{IDR\ low\text{-}high}$ orange bars). Positive values indicate that a given amino acid is fractionally enriched in $IDR_{low\ pLDDT}$ whereas negative values indicate a fractional enrichment in either ordered regions ($\Delta_{IDR\text{-}order}$) or $IDR_{high\ pLDDT}$ regions ($\Delta_{IDR\ low\text{-}high}$). (**B**) The difference between $\Delta_{IDR\text{-}order}$ and $\Delta_{IDR\ low\text{-}high}$ reports on the relative difference in amino-acid usage between ordered regions and $IDR_{high\ pLDDT}$ regions as compared to $IDR_{low\ pLDDT}$ regions. Positive values reflect an increased usage of a given amino acid in $IDR_{high\ pLDDT}$ regions whereas negative values reflect enrichment in ordered regions with respect to $IDR_{low\ pLDDT}$ regions, which we assume reflect canonical disordered regions. (**C**) BLASTp results from querying amino-acid sequences in the PDB (**Methods**) for predicted IDRs in the AFDB that are longer than 10 residues. Percentage of predicted IDRs (hits/total) that were identified in the PDB as a function of the *E*-value and the pLDDT score, with < 50 in orange, ≥ 70 in cyan, and ≥ 90 in blue. Box plots of the number of aligned sequences (**D**), average alignment depth (**E**), average positional conservation (**F**). Panel D is not significant whereas E and F have *p*-values (Mann-Whitney) < 0.0001 when comparing pLDDT < 50 and the other groups.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

IDRs with high pLDDT scores have limited similarity to PDB sequences but are more positionally conserved than IDRs with low pLDDT scores

Next, we searched the PDB for IDRs with considerable sequence similarity, defined as regions that have over 30% sequence identity over 60% sequence coverage. We hypothesized that there should be more $IDR_{high\ pLDDT}$ sequences with similarity to sequences in the PDB than $IDR_{low\ pLDDT}$ sequences. An enrichment in similarity for $IDR_{high\ pLDDT}$ sequences could indicate that AlphaFold2 is matching template structures of these IDRs that were used in training. Indeed, we found that $IDR_{high\ pLDDT}$ sequences are significantly enriched over $IDR_{low\ pLDDT}$ sequences for similarity to PDB sequences (**Figure 5C**). The percentage of IDRs with pLDDT scores ≥ 70 that have confident BLASTP hits (E-value < 1e-3 or < 1e-6) in the PDB is more than 3-fold higher than IDRs with low pLDDT scores (**Figure 5C**). However, it is important to note that the percentage of high-quality hits in the PDB relative to the total number of predicted IDRs in each pLDDT threshold is very low, *i.e.*, a maximum hit rate of 4% was obtained (**Figure 5C**). Therefore, AlphaFold2 has not simply templated structures of IDRs from the PDB, as the overall coverage of IDR sequences in the PDB remains below 4%.

Given our above analyses, we asked if the structural predictions for IDRs with high pLDDT scores could reflect higher positional sequence conservation of these IDRs when compared to IDRs with low pLDDT scores. IDRs that conditionally fold have been previously shown to have higher levels of positional amino-acid conservation than IDRs in general (Bellay et al. 2011; Colak et al. 2013). To compute the positional sequence conservation, we constructed MSAs for predicted IDR sequence across different pLDDT categories using homologous sequences retrieved from the ENSEMBL database (Howe et al. 2021). The MSAs contained nearly identical numbers of sequences for each of the three classes of IDRs (pLDDT scores < 50, ≥ 70, and ≥ 90) (**Figure 5D**), yet the average alignment depth was significantly enriched in IDRs with pLDDT scores ≥ 70 and ≥ 90, relative to those with pLDDT scores < 50 (p-value < 0.0001) (**Figure 5E**). Moreover, the quality of the alignments was higher for IDRs with high pLDDT scores compared to those with low pLDDT scores, as evidenced by greater levels on average of positional conservation (**Figure 5F**).

Overall, our sequence analysis of predicted IDRs demonstrates that those with high pLDDT scores have higher sequence similarity to the sequences in the PDB than predicted IDRs with low pLDDT scores. However, the overall coverage of IDR sequences in the PDB remains low, with only 4% of high-scoring IDR

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequences displaying similarity (E-value < 0.001) to PDB sequences. Moreover, IDRs with high pLDDT scores are more positionally conserved, with nearly 60% sequence conservation on average, and contain fewer gaps than predicted IDRs with low pLDDT scores. Given that AlphaFold2 relies on MSAs as input for its structural predictions (Jumper et al. 2021a), these results provide evidence as to why AlphaFold2 is folding IDRs with high pLDDT scores into confident structures. The fact that IDRs with high pLDDT scores only rarely have sequence homologs in the PDB suggests that the dominant forces behind the AlphaFold2 predictions for these IDRs are high-quality MSAs and the underlying amino-acid compositions, and not structural templating.

AlphaFold2 confidently assigns structures for the majority of known IDRs that conditionally fold

Our bioinformatics analyses provide evidence that IDRs with high pLDDT scores have both compositional differences from and higher quality MSAs than IDRs with low pLDDT scores (**Figure 5**). Given that it has been previously demonstrated that IDRs with high levels of positionally conservation are enriched in IDRs that conditionally fold (Bellay et al. 2011; Colak et al. 2013), we sought to gain additional insight into whether the predicted IDRs with high pLDDT scores are generally found to be conditionally folding. To this end, we investigated the per-residue pLDDT scores for proteins in four databases that contain IDRs/IDPs that fold upon binding (Disfani et al. 2012; Dosztányi et al. 2009; Fichó et al. 2017; Schad et al. 2018): the Database of Disordered Binding Sites (DIBS) (Schad et al. 2018), Mutual Folding Induced by Binding (MFIB) (Fichó et al. 2017), molecular recognition feature (MoRF) (Disfani et al. 2012), and ANCHOR (Dosztányi et al. 2009) databases. We filtered these databases for regions of human proteins that mapped to the AFDB (Methods) and were left with a total of *ca.* 14,000 residues for further analysis. Remarkably, the per-residue pLDDT scores for the IDRs in these four databases are significantly higher than the background “IDR-ome”, with AlphaFold2 assigning confident scores to *ca.* 59% of all residues in the databases, or between 30-80% when each database is analyzed separately (**Supplementary Figure 7**). By contrast, for all IDRs, the percentage of residues with confident pLDDT scores is only 14.3%. Therefore, IDR sequences with confident and very confident pLDDT scores seem to be enriched in experimentally validated MoRFs and other IDRs that conditionally fold upon binding or PTM. Overall, this analysis further supports our hypothesis that IDRs/IDPs with confident and very confident pLDDT scores are likely to be conditional folders.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Leveraging AlphaFold2 to discover conditionally folding IDRs in other organisms

Our findings indicate that the pLDDT score of AlphaFold2 can be used in combination with sequence-based disorder predictors to identify conditionally folded IDRs. Given that AlphaFold2 assigns high-confidence pLDDT scores to IDRs that conditionally fold, we sought to leverage this information to discover IDRs that conditionally fold. For example, in humans there are *ca.* 350 IDRs that are known to conditionally fold based on the largest databases, MFIB and DIBS (Fichó et al. 2017; Schad et al. 2018). Our analyses herein have identified over 8,000 IDRs greater than 10 residues in length that have high-confidence pLDDT scores, which represents a 22-fold increase in the number of conditionally folded IDRs. Moreover, we identified that only *ca.* 4% of IDRs with high-confidence AlphaFold2 structures have sequence similarity to sequences in the PDB (excluding NMR sequences) (**Figure 5B**). Now, with access to more than 8,000 AlphaFold2 structures of IDRs that are putative conditional folders, we have significantly expanded the structural coverage of conditionally folding IDRs by a factor of 25.

In the human proteome, roughly 30% of residues are predicted to be disordered, of which nearly 15% are predicted by AlphaFold2 to conditionally fold. Given that the percentage of intrinsic disordered residues in the proteome has increased from archaea to bacteria to eukaryotes (Gao et al. 2021), we wondered how the percentage of conditionally folded IDRs has changed. To this end, we used IUPred2A (Mészáros et al. 2018) to predict the IDRs in other AFDBs that are publicly available, as IUPred2A software program is *ca.* 100-fold faster than SPOT-Disorder and provides a balance between the calculation speed and accuracy of the prediction (Necci et al. 2021). We first compared the predicted number of disordered residues and conditionally folded IDRs in the human proteome, as obtained by using the IUPred2A and SPOT-Disorder predictions to filter the AFDB. We found that IUPred2A and SPOT-Disorder give comparable results: 32% *vs.* 25% of all residues are predicted to be disordered with 14.3% *vs*. 17.6% of predicted disordered residues having pLDDT scores greater than or equal to 70 for SPOT-Disorder and IUPred2A, respectively (**Supplementary Table 1**, **Supplementary Table 2**). Thus, although IUPred2A underestimates the extent of disorder, the boost in calculation speed provides an attractive approach to perform more high-throughput calculations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We extracted the IUPred2A-predicted IDRs from the 23 AFDBs shown in **Figure 6A**, including four archaeal, seven bacterial, and 12 eukaryotic organisms. As expected, the percentage of disordered residues in the proteome increased from archaea to bacteria to eukaryotes, with minimum and maximum values of 1.0% and 28.1% obtained for *M. janaaschii* and *L. infantum*, respectively (**Figure 6A**). Interestingly, the percentage of IDRs with high-confidence pLDDT scores showed an inverse relation with the overall disordered content. That is, organisms with fewer predicted IDRs have a higher proportion of IDRs with high-confidence pLDDT scores. This result suggests that conditionally folded IDRs are the dominant type of IDRs in the archaea and bacteria examined here, where the percentage of IDRs with high-confidence pLDDT scores ranges from 56% (*M. tuberculosis*) to 81.1% (*T. maritima*). By contrast, in the eukaryotes analyzed here, conditionally folded IDRs appear to be the minority, with minimum and maximum values of 6.6% (*P. falciparum*) and 21.1% (*S. pombe*) (**Figure 6A**). The inverse relation between the percentage of disordered residues in the proteome and the percentage of IDRs that conditionally fold suggests that an upper bound of *ca.* 5% of residues in the proteome localize to conditionally folded IDRs (**Figure 6B**).

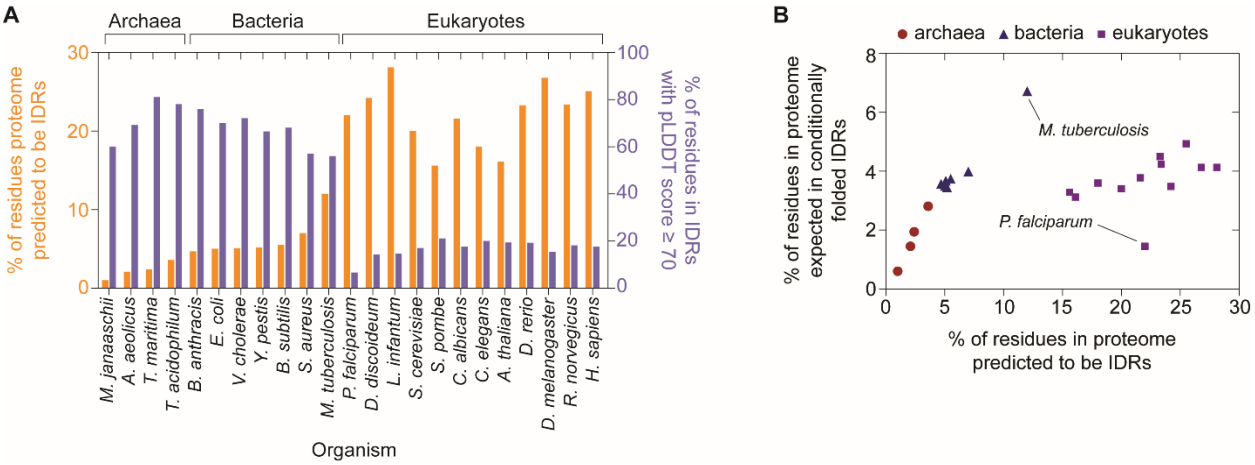

Figure 6. Using AlphaFold2 to discover conditionally folded IDRs in archaea, bacteria, and eukaryotes. (**A**) For each species listed, the percentage of disordered residues in the proteome (predicted by IUPred2A) is shown in orange on the left y-axis. The percentage of disordered residues with pLDDT scores ≥ 70 (*i.e.*, conditionally folded IDRs) is shown in blue on the right y-axis. (**B**) The percentage of residues in the proteome of each organism from panel A that are both predicted to be disordered and to conditionally fold (pLDDT scores ≥ 70).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

The application and development of machine learning methods in structural biology has revolutionized the field of protein structure prediction (AlQuraishi 2021; Baek et al. 2021; Evans et al. 2021; Jumper et al. 2021a). AlphaFold2 can predict the structures of most globular proteins to within experimental accuracy (Jumper et al. 2021a), and the AFDB presently contains structural predictions for *ca.* 98.5% of the human proteome (Tunyasuvunakool et al. 2021). Such remarkable achievements go hand in hand with method developments in structural biology (Eastman et al. 2017; Li et al. 2013; Punjani et al. 2017; Scheres 2012), structural bioinformatics (Andreeva et al. 2020; Mariani et al. 2013; Mistry et al. 2021; Zemla 2003), and protein sequence analysis (Johnson et al. 2010; Remmert et al. 2011), as well as access to the wealth of information stored in publicly available databases, such as the PDB (Bonvin 2021; Burley & Berman 2021; Burley et al. 2019; Goodsell et al. 2020) and a multitude of genomic repositories (Howe et al. 2021). Indeed, either directly or indirectly, AlphaFold2 relied on the methods and databases listed above to predict structures at the proteome level.

However, approximately 30% of the human proteome is intrinsically disordered, with over 60% of all human proteins containing at least one IDR longer than 30 residues in length (Tsang et al. 2020; Van Der Lee et al. 2014). Thus, it is essential to critically analyze the AlphaFold2-predicted structures of IDRs, as such regions cannot be accurately described by a single, static structure. Here, we have shown that there are thousands of predicted IDRs in the human proteome, summing to nearly half a million residues, that are ascribed confident or very confident pLDDT scores in the AFDB (**Figure 1A**), and thus have confidently predicted three-dimensional structures. However, by definition, IDRs do not adopt a stable structure; rather, IDRs rapidly interconvert between conformations and may only exhibit transient secondary structure elements. A more suitable representation of IDRs requires conformational ensembles that describe the free energy landscapes of IDRs, and thus their accessible conformations (Davey 2019; Jensen et al. 2014; Mittag & Forman-Kay 2007; Wei et al. 2016). Conformational ensembles encode all of the structural details of IDRs that, in turn, regulate their interactions with other biomacromolecules and overall functions. Thus, despite high pLDDT scores that suggest high confidence, we caution the users of the AFDB in their interpretation of the static structures for IDRs/IDPs with confident AlphaFold2 structures. The

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

amino-acid sequences of these IDRs/IDPs are consistent with intrinsically disordered sequences, and thus will populate ensembles of disordered conformations that do not autonomously fold.

Despite the obvious problems associated with assigning a static AlphaFold2 structure to IDRs, one valuable metric that emerges from our analyses is an estimate on the fraction of IDR residues that conditionally fold, either upon binding or post-translational modification. We showed that high-confidence AlphaFold2 structures of IDRs/IDPs often can capture a folded conformation that forms in the presence of a specific binding partner or upon post-translational modification (**Figure 2, Figure 3, Figure 4**), even though the structures in the AFDB were predicted in the absence of such binding partners or post-translational modifications. There are presently *ca.* 350 known IDRs that conditionally fold, based on the MFIB and DIBS databases. Thus, the *ca.* 8,000 IDRs with high-confidence AlphaFold2 structures provide a useful resource to interrogate the sequence and structural properties of IDRs that acquire folds during their function. While it will be difficult to predict *a priori* which IDRs will fold following post-translational modification, it may be possible to use high-throughput rigid-body docking to yield insight into the globular domains responsible for folding upon binding for some of these IDRs. Although we note that even small regions of low-confidence AlphaFold2 IDR structures can interfere with the docking accuracy (**Supplementary Figure 4**).

Even though IDPs/IDRs may fold upon binding to a globular domain or upon post-translational modification, the structural plasticity of IDRs and the interconversion between different structural states is essential to their biological function. An example of this is provided by our analysis of IDRs/IDPs that have been experimentally captured in multiple conformations (**Figure 3**). AlphaFold2 predicts that the IDR/IDP would adopt only one of the experimentally observed conformations. If AlphaFold2 can recover the inherent conformational heterogeneity in the bound/modified forms of IDRs/IDPs, such predictions would be valuable to generate testable hypotheses and search for plausible structured conformations that are accessible to a given IDR/IDP. Interestingly, two recent studies proposed that either reducing the depth of the MSA or performing rational *in silico* mutagenesis within the MSA can drive AlphaFold2 to sample alternate conformations (Alamo et al. 2021; Stein & Mchaourab 2021). It remains to be tested if this approach is generally applicable and amenable to IDRs/IDPs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The bias of the traditional structure-function paradigm has often led to an assumption that IDPs and IDRs are disordered since they lack their binding partners. While not quantitatively rigorous, the finding that *ca.* 15% of human IDRs have high-confidence structure predictions suggests that on the order of this number of IDRs might be expected to fold upon binding. At present, we do not know the rate with which AlphaFold2 incorrectly predicts conditionally folded IDRs; however, noting that AlphaFold2 correctly identified *ca.* 60% of IDRs that are known to conditionally fold, we can estimate an upper bound of conditionally folded IDRs at 25% (0.15/0.60). A percentage of conditionally folded human IDRs between 15-25% correlates with previous observations that a minority of human IDRs display significant degrees of positional sequence conservation (Colak et al. 2013), with many of the positionally conserved IDRs identified as those that fold upon binding (Bellay et al. 2011; Colak et al. 2013). Positional conservation is atypical of IDRs that generally evolve rapidly and show low positional conservation, even though there is significant conservation of bulk molecular features (Zarin et al. 2019, 2021). Therefore, there may be purifying selection upon IDR sequences that function by conditional folding, such that the sequence only slowly evolves in order to maintain the overall fold of the bound/modified form of the IDR.

Collectively, these results lead to the hypothesis that if only 15-25% of IDRs are conditionally folded, as suggested by the AFDB and evolutionary analyses, then the majority of IDRs would function in the absence of stable structure, which would include IDRs involved in discrete dynamic or "fuzzy" complexes (Borgia et al. 2018; Fuxreiter 2019; Mittag et al. 2008; Tompa & Fuxreiter 2008) and those with low complexity sequences that participate in dynamic, exchanging condensed phases of biomolecular condensates (Murthy & Fawzi 2020; Peran & Mittag 2020). Our analysis of the percentage of conditionally folded IDRs in various organisms reveals that archaea and bacteria, which have lower disordered content throughout the proteome (Gao et al. 2021), have a relatively higher percentage of IDRs that conditionally fold (**Figure 6**). This suggests that the majority in IDRs in archaea and bacteria will conditionally fold, which means that fuzzy complexes occur less frequently in such organisms. By contrast, eukaryotes appear to have a majority of IDRs that do not acquire stable folds during their function (**Figure 6**).

At the most fundamental level, high-resolution biomolecular structures provide snapshots of the molecular mechanisms that drive all cellular processes and inspire a deeper understanding of how biology happens. At a more practical level, biomolecular structures enable structure-based drug discovery,

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

including the identification of drug-like molecules and the optimization of their binding affinities (Murray & Rees 2009), as well as the prediction of the absorption, distribution, metabolism, excretion and toxicity profile of a drug (Moroy et al. 2012). The PDB has served as an open-access platform for structure-based drug design, evidenced by a recent analysis of 210 drugs that were approved by the US Food and Drug Administration (FDA) between 2010 and 2016: nearly 6,000 structures were available in the PDB for 88% of the drugs and/or targets, with approximately 50% of these structures deposited in the PDB >10 years prior to FDA approval of the drug (Westbrook & Burley 2019). Therefore, it is logical to assume that the increased structural coverage of the (ordered) human proteome afforded by the AFDB (Porta-Pardo et al. 2021) will lead to more structure-based drug discovery. Because IDRs/IDPs are enriched in many signaling pathways (Wright & Dyson 2015), they are highly desirable drug targets (Metallo 2010); however, it is notoriously difficult to target IDRs/IDPs with pharmaceuticals, which likely is due to their absence of stable structures, large interaction networks (Teilum et al. 2021), and interconversion between different structural forms (Metallo 2010). Having access to high-confidence AlphaFold2-predicted structures for the *ca.* 15% of conditionally folded IDRs/IDPs may accelerate the targeting of IDRs/IDPs with rationally designed inhibitors.

While the structural predictions generated by AlphaFold2 will certainly accelerate the pace of biomedical discovery, there remains a huge need for experimental (Bhowmick et al. 2016) and bioinformatic (Zarin et al. 2019, 2021) approaches to address the majority of IDRs that likely function in the absence of folded structure. With increased experimental data on IDRs/IDPs, including integrative structural modelling (Bottaro et al. 2020; Choy & Forman-Kay 2001; Gomes et al. 2020; Krzeminski et al. 2013; Lincoff et al. 2020; Ozenne et al. 2012; Salmon et al. 2010), machine-learning methods promise to provide new insights into disordered protein conformational states and functional mechanisms (Lindorff-Larsen & Kragelund 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Sequence-based prediction of IDRs in the human proteome

We obtained all protein sequences from the human proteome from the UniProt database (reference proteome number UP000005640, downloaded in November 2021). This reference human proteome contains 20,959 unique UniProt IDs that have a total of 11,472,924 residues. To identify IDRs in the human proteome, we used SPOT-Disorder (Hanson et al. 2017), which was recently identified as one of the most accurate predictors of disorder (Necci et al. 2021) and gave the closest agreement with experimentally determined intrinsic disorder based on NMR data (Dass et al. 2020; Nielsen & Mulder 2019). Regions of the proteome that were not predicted to be disordered were assumed to be ordered. For analysis of the per-residue pLDDT scores, the SPOT-Disorder predictions were used without filtering for consecutive residue length (**Figure 1A**); in the bioinformatic analyses of **Figure 5**, to exclude very short segments, we filtered the SPOT-Disorder predictions to include only the regions with predicted consecutive disorder greater than 10 residues.

For the analysis of sequence-based predictors of disorder in **Figure 2A**, we used the software packages metapredict, SPOT-Disorder, DISOPRED3, and IUPred2A (Emenecker et al. 2021; Hanson et al. 2017; Jones & Cozzetto 2015; Mészáros et al. 2018). SPOT-Disorder was run as noted above. The webserver versions of the other three software programs were used with default parameters.

Extraction of per-residue pLDDT scores from the AFDB

Per-residue pLDDT scores were extracted from each PDB file in the database using an in-house Python script. The per-residue SPOT-Disorder predictions of disorder were then mapped onto the AFDB in order to split the AlphaFold2 data into predicted regions of disorder and order. In this way, our analysis should not be biased by the assignment of secondary structure in the AlphaFold2-generated structures; instead, we relied on SPOT-Disorder to identify the IDRs and ordered regions. The SPOT-Disorder-predicted regions of disorder were further split into the AlphaFold2-designated levels of confidence: very low (≤ 50), low (≤ 70), confident (≥ 70 x < 90), or very confident (≥ 90) pLDDT scores. The Python script as well as the output file that contains only the UniProt ID, the pLDDT scores, and the residue types and numbers will be available on GitHub.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the organisms listed in **Figure 6**, AFDBs were downloaded from the AlphaFold website in January 2022. For organisms that did not have pre-compiled AFDBs at that point, but did have AlphaFold2 structures available, an in-house Python script was used to automatically download all structures that matched query UniProt IDs from a UniProt proteome file.

PDB files within the AFDB

The repository of human protein structures was downloaded in November 2021 from the AlphaFold Protein Structure Database using the reference proteome number UP000005640. This corresponded to version 1 of the human AFDB, as indicated with the "v1" string in the name of the downloaded file. The database contained 23,391 predicted structures that map to 20,504 unique UniProt IDs, which corresponds to 97.8% of the proteome (containing 20,959 unique entries). The total number of residues in the AFDB is 10,825,508, which is 94.4% of all residues in the proteome (11,472,924 residues) and *ca.* 2.6% more than what was reported in the original work (Tunyasuvunakool et al. 2021), likely reflecting the *ca.* 1% increase in the number of UniProt IDs (20,504 in AFDB *vs.* 20,296 sequences in the original work).

In the AFDB, the greater number of structures (23,391) relative to the number of unique UniProt IDs (20,504) is due to the 2,700-residue limit for AlphaFold2 structures: proteins longer than this threshold are segmented into multiple structures that contain overlapping 1,400-residue fragments (*e.g.*, residues 1-1400, 201-1600, etc.). The human proteome contains 210 proteins longer than 2,700 residues; searching the AFDB for the UniProt IDs that map to these proteins reveals 3,095 AlphaFold2 structures, accounting for the difference in the number of PDB files and unique UniProt IDs.

However, for proteins longer than 2,700 residues, even though the sequences in the multiple PDB files correspond to overlapping 1,400-residue fragments, the residue numbers in the PDB files themselves always begin at number one and end at 1,400 (or some value less than 1,400 if the final segment has fewer residues). Unless accounted for, the residue numbering in these PDB files could cause erroneous mapping of pLDDT scores that are filtered by the SPOT-Disorder predictions, as well as a failure to extract pLDDT scores from residues beyond 1,400 in the sequence. For example, for the 210 proteins that are longer than 2,700 residues, there are 85,846 residues that are predicted to be disordered. These residues will not be

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correctly mapped from SPOT-Disorder to the AFDB unless the residue numbers in the segmented PDB files have been corrected.

Biological Magnetic Resonance Bank

Assigned NMR chemical shifts for the IDRs/IDPs α-synuclein, 4E-BP2, and ACTR were downloaded from the BMRB (Ulrich et al. 2008) using the following entry identification numbers: 6968 (α-synuclein) (Bermel et al. 2006), 5744 (α-synuclein bound to SDS micelles) (Chandra et al. 2003), 19114 (4E-BP2) (Lukhele et al. 2013), 19905 (phosphorylated 4E-BP2) (Bah et al. 2015), 15397 (ACTR) (Ebert et al. 2008), and 5228 (ACTR bound to CBP) (Demarest et al. 2002). All of the entries contained assignments for ^{1}HN, ^{1}Hα, ^{15}N, ^{13}CO, ^{13}Cα, and ^{13}Cβ chemical shifts, except for 5744 (no ^{1}Hα and ^{13}CO assignments) and 15397 (no ^{1}Hα assignments).

Calculation of NMR chemical shifts from an input PDB structure

To simulate the NMR chemical shifts of AlphaFold2-generated structure predictions, we used the SPARTA+ software package (Shen & Bax 2010). Protons were first added to each PDB structure using DYNAMO version 7.2 available via the PDB Utility Web Servers from the Bax Laboratory. The proton-containing PDB files were then uploaded to the SPARTA+ Web Server using default parameters. The backbone and $^{13}C^{\beta}$ chemical shifts were extracted from the output file with an in-house Python script that will be available on GitHub.

Calculation of secondary structure propensity from NMR chemical shifts

We used the SSP software program (Marsh et al. 2006) via the NMRbox (Maciejewski et al. 2017) to calculate the secondary structure propensity (SSP) of the IDRs/IDPs shown in **Figure 2**. We included $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts as input, as recommended in the original SSP publication. Prior to analysis, we re-referenced all of the chemical shifts using standard protocols (Marsh et al. 2006). This is important because secondary chemical shifts, and therefore SSP, are highly sensitive to the internal referencing of the measured chemical shifts, and any errors in referencing will impact the downstream SSP analysis. The SSP-derived re-

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

referencing offset ppm in the ^{13}C dimension for each dataset was -0.390 (BMRB: 6968), -0.483 (5744), +0.373 (191141), +0.106 (19905), -0.051 (15397), and -0.139 (5228).

Sequence similarity between IDRs and the PDB

We ran BLASTP (Altschul et al. 1990) to determine to which extent IDR sequences with different pLDDT scores overlap with sequences in the PDB. To filter the sequences that are in the PDB, we removed duplicates of identical sequences and highly homologues sequences using PISCES (Wang & Dunbrack 2003) with the following parameters: maximum pairwise percent sequence identity (75%), resolution (0-3.5A), minimum chain length (40), and the maximum chain length (10,000). The sequence data included all X-ray structures that met the aforementioned criteria but excluded all NMR and cryo-EM entries as well as sequences that map to regions of X-ray structures with no electron density. We separately downloaded all X-ray structures that contain 40 or fewer residues. The curation resulted in a set of *ca.* 4,000 unique sequence entries, which were then used to create the BLAST database. BLASTP was run with E-value cut-off values of 1e-3 and 1e-6, with restrictions on sequence identity (>30%) and coverage criteria (>60%).

Evaluation of positional sequence conservation in IDRs

Positional sequence conservation was computed for alignments of IDRs that were distributed in three sets with different cut-offs of pLDDT scores (see "Mapping pLDDT scores to IDRs"). Only IDR sequences with 10 or more residues with consecutive pLDDT score below or above the desired threshold were considered. To compute positional conservation across MSA columns, we used a modified metric of Shannon's entropy, the so-called property entropy as introduced by Capra and Singh (Capra & Singh 2007).

Bioinformatic analysis of the predicted IDRs in the AFDB

There are 10,825,508 residues in the AFDB, of which 3,539,799 are predicted by SPOT-Disorder to be disordered (**Supplementary Table 1**), 7,127,685 to be ordered, and 158,024 are not mapped. The latter is consistent with the *ca.* 98.5% coverage of the human proteome in the AFDB (Tunyasuvunakool et al. 2021). From these numbers, we can also calculate the percentage of residues in SPOT-Disorder-predicted IDRs, which amounts to 32.7% of the AFDB, in agreement with literature values (Necci et al. 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, the SPOT-Disorder-predicted IDRs were further split into those with low pLDDT scores (< 70) and those with high pLDDT scores (≥ 70). A total of 506,101 residues are in $IDR_{high\ pLDDT}$ (pLDDT ≥ 70) as compared to 3,033,698 in $IDR_{low\ pLDDT}$ (pLDDT < 70). Using an in-house Python script, we calculated the amino-acid frequencies in each of the following categories: ordered regions, IDRs with low pLDDT scores ($IDR_{low\ pLDDT}$), and IDRs with high pLDDT scores ($IDR_{high\ pLDDT}$). The amino-acid frequencies were normalized to the total number of amino acids in each category to yield the percentage of the total:

$$AA_{\mathrm{freq,i,j}} = AA_{i,j} \Big/ \sum_{i=1}^{20} AA_{i,j} \qquad (1)$$

Where $AA_{i,j}$ stands for the number of amino acids of residue type *i*, with the indices *i* and *j* respectively indicating the amino acid type (A, C, D, … , V, W, Y) and the category of residues analyzed (ordered, $IDR_{low\ pLDDT}$, $IDR_{high\ pLDDT}$). The summation in the denominator of eq 1 refers to the total number of amino acids in each category.

Mean net charge and mean hydrophobicity of $IDR_{low\ pLDDT}$ and $IDR_{high\ pLDDT}$ sequences were calculated according to Uversky *et al.* (Uversky et al. 2000) using an in-house Python script. Briefly, the net charge of a given IDR/IDP sequence was computed at pH 7. The pK_a of histidine residues was set to 6.5 to match an experimental determination of the His pK_a in an IDP recorded in the presence of physiological salt concentrations (Croke et al. 2011). The absolute value of the net charge of the IDR/IDP was then divided by the total number of residues to obtain the mean net charge. The mean hydrophobicity of an IDR was computed using a normalized version of the Kyte-Doolittle hydropathy scale, such that the values ranged between 0 and 1. The hydropathy value of each residue in the IDR/IDP was then averaged over a sliding window of five residues. The mean hydropathy was finally obtained by computing the sum of all hydropathy values (via the sliding window) and then dividing by the total number of residues.

Databases of IDRs/IDPs that fold upon binding

The Disordered Binding Sites (DIBS) (Schad et al. 2018) and Mutual Folding Induced Binding (MFIB) (Fichó et al. 2017) databases were accessed on December 2, 2021. The versions of these databases are 08-05-2019 and 26-06-2017, respectively, and contain 772 and 205 entries of which there are 501 and 246 unique IDPs/IDRs that are folded while bound to a globular domain or another IDP/IDR. We extracted the sequence

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regions of the IDRs/IDPs in each database and searched for the corresponding regions in the human AFDB. In total, there were 265 (6878) and 72 (6365) IDPs/IDRs (residues) for further examination. As anticipated, these IDRs/IDPs are significantly enriched in confident and high confident pLDDT scores: DIBS has 29% of residues with pLDDT scores ≥ 70 and 8.7% ≥ 90, while MFIB has 81.9% ≥ 70 and 62.7% ≥ 90. The percentage of residues with confident (≥ 70) pLDDT scores is nearly 2- and 5-fold enriched relative to the entire SPOT-Disorder-predicted "IDR-ome" for DIBS and MFIB, respectively.

This analysis was also performed on an additional database of short IDRs that fold upon binding (Disfani et al. 2012). Disorder-to-order transitions upon binding commonly occur in molecular recognition features (MoRFs), which are short disordered segments (5-25 residues) that exist within longer IDRs that bind to a globular domain (Disfani et al. 2012). For the 61 MoRFs from human IDRs, totalling 949 residues, that exist in the AFDB, we find that the fraction of confident (35.9%) and very confident (13.3%) per-residue pLDDT scores is respectively nearly two- and three-fold higher than the values for all SPOTD-predicted IDRs (**Supplementary Figure 7**). Furthermore, we observed an additional increase in the fractions of confident and very confident pLDDT scores when analyzing only the MoRFs that fold upon binding into α-helix or β-strand secondary structure elements, while removing MoRFs that remain as coils when bound. This filtering procedure leaves 19 proteins and 269 residues in total, with the fraction of confident (*ca.* 59.9%) and very confident (19.3%) pLDDT scores *ca.* 4-fold higher than the corresponding values for all SPOT-Disorder-predicted IDRs

Simulation of biophysical parameters from an input PDB structure

As outlined in Supplementary Figure 3 and the Supplementary Appendix, biophysical experiments can be performed and compared to predictions derived from the AFDB structural model. Such comparisons would be able to rapidly report on the accuracy of the model relative to the conformations sampled in solution.

Circular dichroism (CD): CD spectra of proteins are sensitive to global secondary structure content and do not require much sample. The webserver PDB2CD (Mavridis & Janes 2017) was used with default parameters to simulate the CD spectrum of the AFDB structure of human α-synuclein and to compare to experimental data (Fusco et al. 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Translational diffusion (PFG-NMR): The software package HYDROPRO was used to calculate hydrodynamic properties of α-synuclein based on its structure in the AFDB (Ortega et al. 2011) (**Supplementary Figure 3B**). Unless otherwise specified, default parameters were used (*e.g.,* a non-overlapping shell model was used with shell model set to the atomic level and a 2.84-Å radius of atomic elements). The temperature was set to 15 °C to match the experimental conditions (Ramanujam et al. 2020), and the solvent viscosity was adjusted accordingly to 1.1366 cP based on the value reported by NIST. The AlphaFold2 structure from the AFDB was then loaded and the calculation was run. The predicted translational diffusion coefficient from HYDROPRO is 5.141 x 10^{-11} m^2 s^{-1}, which is approximately 10% smaller than the experimentally measured value of 5.71 ± 0.02 x 10^{-11} m^2 s^{-1} (Ramanujam et al. 2020), suggesting that the AlphaFold2 structure is more extended than the conformation of α-synuclein observed experimentally. To generate a simulated plot of signal decay (I_j / I_0) caused by translational diffusion during a BPP-LED pulse sequence, equation 2 below was used:

$$I_j = I_0 \, e^{-\gamma^2 G_j^2 \delta^2 \left(\Delta - \frac{\delta}{3} - \frac{\tau}{2}\right) D} \qquad (2)$$

Where the measured signal intensity, I_j, depends on the exponential term above that contains the square strength of the applied gradient (G_j^2), which is varied during the experiment, and a linear contribution from the translational diffusion coefficient (D). The other parameters are either held fixed in the experiment (Δ, the delay time for translational diffusion; δ, the total duration of the encoding gradients; τ, the gradient recovery duration) or are physical parameters (γ, the gyromagnetic ratio of ^{1}H). Values of 267,522,187.44 rad s^{-1} T^{-1}, 200 ms, 3 ms, 200 μs, and 0.668 T m^{-1} were used for γ, Δ, δ, τ, and G_{max}, respectively, to match experimental conditions (Ramanujam et al. 2020). The values of G_j were varied over a range of G_j / G_{max} from 0 to 1.

Small-angle X-ray scattering (SAXS): The software program Crysol (Svergun et al. 1995) was used to simulate SAXS data of human α-synuclein based on the AFDB structure (**Supplementary Figure 3C**). Experimental SAXS data are available for comparison (Ahmed et al. 2021). The software package ATSAS (Manalastas-Cantos et al. 2021) was used to analyze the SAXS data to obtain the radius of gyration (R_g)

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and the maximum distance (D_{max}). The fitted experimental values of R_g and D_{max} are 35.6 ± 0.2 Å and 109 Å, and those derived from the simulated data from the AFDB structure are 42.6 Å and 152 Å, respectively.

Other NMR parameters: The ^{13}Cα chemical shifts in **Supplementary Figure 3D** were simulated with SPARTA+ (Shen & Bax 2010) after protons had been added to the AFDB structure of human α-synuclein and neighbor-corrected random coil chemical shifts were obtained from the SPARTA+ output file. The measured ^{13}Cα chemical shifts were extracted from (Bermel et al. 2006). The $^3J_{HNHα}$ coupling constants in **Supplementary Figure 3E** were simulated from the AFDB structure of human α-synuclein in which protons had been added, as described above. The parameterized form of the Karplus equation used to relate the dihedral angle Φ to the $^3J_{HNHα}$ coupling constant is shown below and based on (Vögeli et al. 2007):

$$^3J_{\mathrm{HNH\alpha}} = 7.97\cos\theta^2 - 1.26\cos\theta + 0.63 \qquad (3)$$

Where θ is the dihedral angle Φ minus 60°, which is then converted to radians. Dihedral angles were computed from the AFDB structure of human α-synuclein using an in-house Python script that uses the BioPython package (Cock et al. 2009). The simulated values of $^3J_{HNHα}$ were compared to those measured experimentally (Mantsyzov et al. 2015). Finally, ^{1}Hα solvent paramagnetic relaxation enhancements (sPREs) were obtained from (Hartlmüller et al. 2019). Simulated sPREs were performed using the sPRE-calc software program (Gong et al. 2017), and the simulated rates were then scaled to roughly match the lowest experimentally reported sPRE values.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We thank William Ford Freyberg (University of Wisconsin-Madison, USA) for critical comments on the manuscript, Dr. Giuliana Fusco (University of Cambridge, UK) for sharing the experimental CD spectrum of α-synuclein, Dr. Kresten Lindorff-Larsen (University of Copenhagen, Denmark) for making the α-synuclein SAXS data available via GitHub, and Emil Spritzer and Dr. Tobias Madl (Medical University of Graz, Austria) for sharing the solvent PRE data from α-synuclein. This study made use of NMRbox: National Center for Biomolecular NMR Data Processing and Analysis, a Biomedical Technology Research Resource (BTRR), which is supported by NIH grant P41GM111135 (NIGMS). TRA and IP were supported by a Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research (CIHR) and a LiUNA! Fellowship for Research Innovation from The Hospital for Sick Children, respectively. AMM and JDF-K acknowledge support from the CIHR (CIHR Foundation Grant (grant no. FDN-148375) to JDF-K; CIHR grant no. PJT-148532 to AMM and JDF-K) and the Canada Foundation for Innovation (CFI) for funding to AMM. JDF-K holds a Canada Research Chair in Intrinsically Disordered Proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Ahmed MC, Skaanning LK, Jussupow A, Newcombe EA, Kragelund BB, et al. 2021. Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Front. Mol. Biosci.* 8:

Alamo D del, Sala D, Mchaourab HS, Meiler J. 2021. Sampling the conformational landscapes of transporters and receptors with AlphaFold2. *bioRxiv*. 2021.11.22.469536

Alberti S, Dormann D. 2019. Liquid-Liquid Phase Separation in Disease. *Annu. Rev. Genet.* 53:171–94

Alderson TR, Lee JH, Charlier C, Ying J, Bax A. 2018. Propensity for cis-Proline Formation in Unfolded Proteins. *ChemBioChem*. 19(1):

AlQuraishi M. 2021. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* 65:1–8

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215(3):403–10

Andreeva A, Kulesha E, Gough J, Murzin AG. 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48(D1):D376–82

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science (80-.).* 181(4096):223–30

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science (80-.).* 373(6557):871–76

Bah A, Vernon RM, Siddiqui Z, Krzeminski M, Muhandiram R, et al. 2015. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*. 519(7541):106–9

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. *Science (80-.).* 294(5540):93–96

Baker JMR, Hudson RP, Kanelis V, Choy WY, Thibodeau PH, et al. 2007. CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nat. Struct. Mol. Biol. 2007 148*. 14(8):738–45

Barnes CA, Shen Y, Ying J, Takagi Y, Torchia DA, et al. 2019. Remarkable Rigidity of the Single α-Helical Domain of Myosin-VI As Revealed by NMR Spectroscopy. *J. Am. Chem. Soc.* 141(22):9004–17

Bellay J, Han S, Michaut M, Kim TH, Costanzo M, et al. 2011. Bringing order to protein disorder through

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comparative genomics and genetic interactions. *Genome Biol.* 12(2):1–15

Bermel W, Bertini I, Felli IC, Lee YM, Luchinat C, Pierattelli R. 2006. Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* 128(12):3918–19

Bertoncini CW, Jung YS, Fernandez CO, Hoyer W, Griesinger C, et al. 2005. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc. Natl. Acad. Sci. U. S. A.* 102(5):1430–35

Bhowmick A, Brookes DH, Yost SR, Dyson HJ, Forman-Kay JD, et al. 2016. Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc.* 138(31):9730–42

Bodner CR, Dobson CM, Bax A. 2009. Multiple tight phospholipid-binding modes of alpha-synuclein revealed by solution NMR spectroscopy. *J. Mol. Biol.* 390(4):775–90

Bonvin AMJJ. 2021. 50 years of PBD: a catalyst in structural biology. *Nat. Methods*. 18(5):448–49

Borcherds W, Bremer A, Borgia MB, Mittag T. 2021. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr. Opin. Struct. Biol.* 67:41–50

Borcherds W, Theillet FX, Katzer A, Finzel A, Mishall KM, et al. 2014. Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* 10(12):1000–1002

Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, et al. 2018. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*. 555(7694):61–66

Bottaro S, Bengtsen T, Lindorff-Larsen K. 2020. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *Methods Mol. Biol.* 2112:219–40

Bozoky Z, Krzeminski M, Chong PA, Forman-Kay JD. 2013a. Structural changes of CFTR R region upon phosphorylation: a plastic platform for intramolecular and intermolecular interactions. *FEBS J.* 280(18):4407–16

Bozoky Z, Krzeminski M, Muhandiram R, Birtley JR, Al-Zahrani A, et al. 2013b. Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions. *Proc. Natl. Acad. Sci. U. S. A.* 110(47):E4427–36

Breidenbach MA, Brunger AT. 2004. Substrate recognition strategy for botulinum neurotoxin serotype A. *Nature*. 432(7019):925–29

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Burley SK, Berman HM. 2021. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. *Structure*. 29(6):515–20

Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47(D1):D464–74

Camilloni C, De Simone A, Vranken WF, Vendruscolo M. 2012. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*. 51(11):2224–31

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics*. 23(15):1875–82

Chandra S, Chen X, Rizo J, Jahn R, Südhof TC. 2003. A broken alpha-helix in folded alpha-Synuclein. *J. Biol. Chem.* 278(17):15313–18

Chen X, Tomchick DR, Kovrigin E, Araç D, Machius M, et al. 2002. Three-Dimensional Structure of the Complexin/SNARE Complex. *Neuron*. 33(3):397–409

Choy WY, Forman-Kay JD. 2001. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* 308(5):1011–32

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25(11):1422–23

Colak R, Kim TH, Michaut M, Sun M, Irimia M, et al. 2013. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput. Biol.* 9(4):

Conicella AE, Dignon GL, Zerze GH, Schmidt HB, D'Ordine AM, et al. 2020. TDP-43 α-helical structure tunes liquid-liquid phase separation and function. *Proc. Natl. Acad. Sci. U. S. A.* 117(11):5883–94

Cornilescu G, Delaglio F, Bax A. 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR 1999 133*. 13(3):289–302

Croke RL, Patil SM, Quevreaux J, Kendall DA, Alexandrescu AT. 2011. NMR determination of pKa values in α-synuclein. *Protein Sci.* 20(2):256–69

Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu R V. 2016. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U. S. A.* 113(20):5616–

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Das RK, Pappu R V. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.* 110(33):13392–97

Dass R, Mulder FAA, Nielsen JT. 2020. ODiNPred: comprehensive prediction of protein order and disorder. *Sci. Rep.* 10(1):

Davey NE. 2019. The functional importance of structure in unstructured protein regions. *Curr. Opin. Struct. Biol.* 56:155–63

Dawson JE, Bah A, Zhang Z, Vernon RM, Lin H, et al. 2020. Non-cooperative 4E-BP2 folding with exchange between eIF4E-binding and binding-incompatible states tunes cap-dependent translation inhibition. *Nat. Commun.* 11(1):

Demarest SJ, Martinez-Yamout M, Chung J, Chen H, Xu W, et al. 2002. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*. 415(6871):549–53

Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, et al. 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*. 28(12):i75–83

Dosztányi Z, Mészáros B, Simon I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*. 25(20):2745

Dyson HJ, Wright PE. 2021. NMR illuminates intrinsic disorder. *Curr. Opin. Struct. Biol.* 70:44–52

Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, et al. 2017. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13(7):e1005659

Ebert MO, Bae SH, Dyson HJ, Wright PE. 2008. NMR relaxation study of the complex formed between CBP and the activation domain of the nuclear hormone receptor coactivator ACTR. *Biochemistry*. 47(5):1299–1308

Eliezer D. 2007. Characterizing residual structure in disordered protein States using nuclear magnetic resonance. *Methods Mol. Biol.* 350:49–67

Eliezer D, Kutluay E, Bussell R, Browne G. 2001. Conformational properties of alpha-synuclein in its free and lipid-associated states. *J. Mol. Biol.* 307(4):1061–73

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Emenecker RJ, Griffith D, Holehouse AS. 2021. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* 120(20):4312–19

Escobedo A, Topal B, Kunze MBA, Aranda J, Chiesa G, et al. 2019. Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. *Nat. Commun. 2019 101*. 10(1):1–11

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, et al. 2021. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. 2021.10.04.463034

Fichó E, Reményi I, Simon I, Mészáros B. 2017. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics*. 33(22):3682–84

Fusco G, Pape T, Stephens AD, Mahou P, Costa AR, et al. 2016. Structural basis of synaptic vesicle assembly promoted by α-synuclein. *Nat. Commun. 2016 71*. 7(1):1–12

Fuxreiter M. 2019. Fold or not to fold upon binding - does it really matter? *Curr. Opin. Struct. Biol.* 54:19–25

Gao C, Ma C, Wang H, Zhong H, Zang J, et al. 2021. Intrinsic disorder in protein domains contributes to both organism complexity and clade-specific functions. *Sci. Rep.* 11(1):1–18

Gomes GNW, Krzeminski M, Namini A, Martin EW, Mittag T, et al. 2020. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.* 142(37):15697–710

Gong Z, Gu XH, Guo DC, Wang J, Tang C. 2017. Protein Structural Ensembles Visualized by Solvent Paramagnetic Relaxation Enhancement. *Angew. Chemie*. 56(4):1002–6

Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, et al. 2020. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.* 29(1):52–65

Hanson J, Yang Y, Paliwal K, Zhou Y. 2017. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*. 33(5):685–92

Hartlmüller C, Spreitzer E, Göbl C, Falsone F, Madl T. 2019. NMR characterization of solvent accessibility and transient structure in intrinsically disordered proteins. *J. Biomol. NMR*. 73(6–7):305–17

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, et al. 2021. Ensembl 2021. *Nucleic Acids Res.* 49(D1):D884–91

Iešmantavičius V, Dogan J, Jemth P, Teilum K, Kjaergaard M. 2014. Helical propensity in an intrinsically

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disordered protein accelerates ligand binding. *Angew. Chem. Int. Ed. Engl.* 53(6):1548–51

Jensen MR, Zweckstetter M, Huang JR, Blackledge M. 2014. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.* 114(13):6632–60

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. 11:431

Jones DT, Cozzetto D. 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 31(6):857–63

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021a. Highly accurate protein structure prediction with AlphaFold. *Nature*. 596(7873):583–89

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021b. Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinforma.*

Kakeshpour T, Ramanujam V, Barnes CA, Shen Y, Ying J, Bax A. 2021. A lowly populated, transient β-sheet structure in monomeric Aβ 1-42 identified by multinuclear NMR of chemical denaturation. *Biophys. Chem.* 270:

Kim TH, Payliss BJ, Nosella ML, Lee ITW, Toyama Y, et al. 2021. Interaction hot spots for phase separation revealed by NMR studies of a CAPRIN1 condensed phase. *Proc. Natl. Acad. Sci. U. S. A.* 118(23):

Konrat R. 2014. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* 241(1):74–85

Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD. 2013. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*. 29(3):398–99

Kuhlman B, Bradley P. 2019. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* 20(11):681–97

Lazar T, Martínez-Pérez E, Quaglia F, Hatos A, Chemes LB, et al. 2021. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* 49(D1):D404–11

Li X, Mooney P, Zheng S, Booth CR, Braunfeld MB, et al. 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods 2013 106*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10(6):584–90

Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes GNW, et al. 2020. Extended Experimental Inferential Structure Determination Method in Determining the Structural Ensembles of Disordered Protein States. *Commun. Chem.* 3(1):

Lindorff-Larsen K, Kragelund BB. 2021. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *J. Mol. Biol.* 433(20):

Lukhele S, Bah A, Lin H, Sonenberg N, Forman-Kay JD. 2013. Interaction of the eukaryotic initiation factor 4E with 4E-BP2 at a dynamic bipartite interface. *Structure*. 21(12):2186–96

Maciejewski MW, Schuyler AD, Gryk MR, Moraru II, Romero PR, et al. 2017. NMRbox: A Resource for Biomolecular NMR Computation. *Biophys. J.* 112(8):1529–34

Malki A, Teulon J-M, Camacho Zarco A, Chen SW, Adamski W, et al. 2021. Intrinsically Disordered Tardigrade Proteins Self-Assemble into Fibrous Gels in Response to Environmental Stress. *Angew. Chem. Int. Ed. Engl.*

Maltsev AS, Ying J, Bax A. 2012. Impact of N-terminal acetylation of α-synuclein on its random coil and lipid binding properties. *Biochemistry*. 51(25):5004–13

Manalastas-Cantos K, Konarev P V., Hajizadeh NR, Kikhney AG, Petoukhov M V., et al. 2021. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* 54(Pt 1):343–55

Mantsyzov AB, Maltsev AS, Ying J, Shen Y, Hummer G, Bax A. 2014. A maximum entropy approach to the study of residue-specific backbone angle distributions in α-synuclein, an intrinsically disordered protein. *Protein Sci.* 23(9):1275–90

Mantsyzov AB, Shen Y, Lee JH, Hummer G, Bax A. 2015. MERA: A webserver for evaluating backbone torsion angle distributions in dynamic and disordered proteins from NMR data. *J. Biomol. NMR*. 63(1):85–95

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

Marqusee S, Baldwin RL. 1987. Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

design. *Proc. Natl. Acad. Sci. U. S. A.* 84(24):8898–8902

Marsh JA, Dancheck B, Ragusa MJ, Allaire M, Forman-Kay JD, Peti W. 2010. Structural Diversity in Free and Bound States of Intrinsically Disordered Protein Phosphatase 1 Regulators. *Structure*. 18(9):1094–1103

Marsh JA, Singh VK, Jia Z, Forman-Kay JD. 2006. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci.* 15(12):2795–2804

Martin EW, Holehouse AS. 2020. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg. Top. Life Sci.* 4(3):307–29

Mavridis L, Janes RW. 2017. PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates. *Bioinformatics*. 33(1):56–63

Mészáros B, Erdös G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46(W1):W329–37

Metallo SJ. 2010. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* 14(4):481–88

Milles S, Mercadante D, Aramburu IV, Jensen MR, Banterle N, et al. 2015. Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors. *Cell*. 163(3):734–45

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2021. ColabFold - Making protein folding accessible to all. *bioRxiv*. 2021.08.15.456425

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49(D1):D412–19

Mittag T, Forman-Kay JD. 2007. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* 17(1):3–14

Mittag T, Orlicky S, Choy WY, Tang X, Lin H, et al. 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. U. S. A.* 105(46):17772–77

Moroy G, Martiny VY, Vayer P, Villoutreix BO, Miteva MA. 2012. Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discov. Today*. 17(1–2):44–55

Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. 2018. Critical assessment of methods of

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein structure prediction (CASP)—Round XII. *Proteins Struct. Funct. Bioinforma.* 86:7–15

Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinforma.* 23(3):ii–iv

Murray CW, Rees DC. 2009. The rise of fragment-based drug discovery. *Nat. Chem. 2009 13*. 1(3):187–92

Murthy AC, Fawzi NL. 2020. The (un)structural biology of biomolecular liquid-liquid phase separation using NMR spectroscopy. *J. Biol. Chem.* 295(8):2375–84

Necci M, Piovesan D, Hoque MT, Walsh I, Iqbal S, et al. 2021. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*. 18(5):472–81

Nguyen Ba AN, Yeh BJ, Van Dyk D, Davidson AR, Andrews BJ, et al. 2012. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal.* 5(215):

Nielsen JT, Mulder FAA. 2019. Quality and bias of protein disorder predictors. *Sci. Rep.* 9(1):

Nielsen JT, Mulder FAA. 2021. CheSPI: chemical shift secondary structure population inference. *J. Biomol. NMR*. 75(6–7):273–91

Ortega A, Amorós D, García De La Torre J. 2011. Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models. *Biophys. J.* 101(4):892

Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, et al. 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*. 28(11):1463–70

Peran I, Mittag T. 2020. Molecular structure in biomolecular condensates. *Curr. Opin. Struct. Biol.* 60:17–26

Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. 2021. The structural coverage of the human proteome before and after AlphaFold. *bioRxiv*. 2021.08.03.454980

Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods 2017 143*. 14(3):290–96

Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, et al. 2021. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.*

Ramanujam V, Alderson TR, Pritišanac I, Ying J, Bax A. 2020. Protein structural changes characterized by

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

high-pressure, pulsed field gradient diffusion NMR spectroscopy. *J. Magn. Reson.* 312:106701

Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*. 9(2):173–75

Robertson AJ, Courtney JM, Shen Y, Ying J, Bax A. 2021. Concordance of X-ray and AlphaFold2 Models of SARS-CoV-2 Main Protease with Residual Dipolar Couplings Measured in Solution. *J. Am. Chem. Soc.* 143(46):19306–10

Ruff KM, Pappu R V. 2021. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* 433(20):167208

Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, et al. 2010. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132(24):8407–18

Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ. 2018. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins*. 86(Suppl Suppl 1):51

Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. 2018. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics*. 34(3):535–37

Scheres SHW. 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180(3):519–30

Schiavo GG, Benfenati F, Poulain B, Rossetto O, De Laureto PP, et al. 1992. Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature*. 359(6398):832–35

Shen Y, Bax A. 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*. 48(1):13–22

Simm D, Kollmar M. 2018. Waggawagga-CLI: A command-line tool for predicting stable single α-helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS One*. 13(2):

Smith NC, Kuravsky M, Shammas SL, Matthews JM. 2021. Binding and folding in transcriptional complexes. *Curr. Opin. Struct. Biol.* 66:156–62

Spera S, Bax A. 1991. Empirical Correlation between Protein Backbone Conformation and Cα and Cβ 13C Nuclear Magnetic Resonance Chemical Shifts. *J. Am. Chem. Soc.* 113(14):5490–92

Stein RA, Mchaourab HS. 2021. Modeling Alternate Conformations with Alphafold2 via Modification of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multiple Sequence Alignment. *bioRxiv*. 2021.11.29.470469

Sugase K, Dyson HJ, Wright PE. 2007. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*. 447(7147):1021–25

Svergun D, Barberato C, Koch MH. 1995. CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* 28(6):768–73

Swanson CJ, Sivaramakrishnan S. 2014. Harnessing the unique structural properties of isolated α-helices. *J. Biol. Chem.* 289(37):25460–67

Teilum K, Olsen JG, Kragelund BB. 2021. On the specificity of protein-protein interactions in the context of disorder. *Biochem. J.* 478(11):2035–50

Theillet FX, Binolfi A, Bekei B, Martorana A, Rose HM, et al. 2016. Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature*. 530(7588):45–50

Tompa P, Fuxreiter M. 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33(1):2–8

Tsang B, Pritišanac I, Scherer SW, Moses AM, Forman-Kay JD. 2020. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature*. 596(7873):590–96

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, et al. 2008. BioMagResBank. *Nucleic Acids Res.* 36(Database issue):

Uversky VN, Gillespie JR, Fink AL. 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? - PubMed. *Proteins*. 41(3):415–27

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37:215–46

Vacic V, Markwick PRL, Oldfield CJ, Zhao X, Haynes C, et al. 2012. Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder. *PLoS Comput. Biol.* 8(10):

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, et al. 2014. Classification of intrinsically disordered regions and proteins

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. 2021. AlphaFold Protein Structure

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.18.481080; this version posted February 18, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*

Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, et al. 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* 42(Database issue):

Vögeli B, Ying J, Grishaev A, Bax A. 2007. Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. *J. Am. Chem. Soc.* 129(30):9377–85

Wang G, Dunbrack RL. 2003. PISCES: a protein sequence culling server. *Bioinformatics*. 19(12):1589–91

Wei G, Xi W, Nussinov R, Ma B. 2016. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* 116(11):6516–51

Westbrook JD, Burley SK. 2019. How Structural Biologists and the Protein Data Bank Contributed to Recent US FDA New Drug Approvals. *Structure*. 27(2):217

Wishart DS, Sykes BD, Richards FM. 1991. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* 222(2):311–33

Wright PE, Dyson HJ. 2009. Linking folding and binding. *Curr. Opin. Struct. Biol.* 19(1):31–38

Wright PE, Dyson HJ. 2015. Intrinsically disordered proteins in cellular signalling and regulation

Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. 2019. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife*. 8:

Zarin T, Strome B, Peng G, Pritišanac I, Forman-Kay JD, Moses AM. 2021. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife*. 10:1–36

Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31(13):3370–74

Zosel F, Mercadante D, Nettels D, Schuler B. 2018. A proline switch explains kinetic heterogeneity in a coupled folding and binding reaction. *Nat. Commun.* 9(1):

Zweckstetter M. 2021. NMR hawk-eyed view of AlphaFold2 structures. *Protein Sci.* 30(11):2333–37