AlphaFold and Implications for Intrinsically Disordered Proteins

Kiersten M. Ruff and Rohit V. Pappu *

Department of Biomedical Engineering and Center for Science & Engineering of Living Systems (CSELS),
Washington University in St. Louis, Campus Box 1097, St. Louis, MO 63130, USA

Correspondence to Rohit V. Pappu: *pappu@wustl.edu (R.V. Pappu)*
https://doi.org/10.1016/j.jmb.2021.167208
Edited by Sheena E. Radford

Abstract

Accurate predictions of the three-dimensional structures of proteins from their amino acid sequences have come of age. AlphaFold, a deep learning-based approach to protein structure prediction, shows remarkable success in independent assessments of prediction accuracy. A significant epoch in structural bioinformatics was the structural annotation of over 98% of protein sequences in the human proteome. Interestingly, many predictions feature regions of very low confidence, and these regions largely overlap with intrinsically disordered regions (IDRs). That over 30% of regions within the proteome are disordered is congruent with estimates that have been made over the past two decades, as intense efforts have been undertaken to generalize the structure–function paradigm to include the importance of conformational heterogeneity and dynamics. With structural annotations from AlphaFold in hand, there is the temptation to draw inferences regarding the "structures" of IDRs and their interactomes. Here, we offer a cautionary note regarding the misinterpretations that might ensue and highlight efforts that provide concrete understanding of sequence-ensemble-function relationships of IDRs. This perspective is intended to emphasize the importance of IDRs in sequence-function relationships (SERs) and to highlight how one might go about extracting quantitative SERs to make sense of how IDRs function.

Introduction

Two recent reports highlight the coming of age of machine learning and artificial intelligence in structural bioinformatics.[1,2] AlphaFold, a state-of-the-art machine learning model from DeepMind, was shown to be highly successful for predicting the three-dimensional structures of proteins from their amino acid sequences.[1] This success, as assessed by adjudicators of the 14th edition of the Critical Assessment of Protein Structure (CASP) experiment, created quite a stir when it was deployed to predict the structures of 98.5% of proteins in the human proteome. Through a portal hosted by the European Bioinformatics Institute (http://alphafold.ebi.ac.uk) one can now access predictions of protein structures made using AlphaFold. In an instant, detailed, atomic-level models of protein structures have become available for most of the protein sequences in the human proteome. This is a remarkable achievement and highlights the giant leaps that are being realized through trained use of curated knowledge bases that provide information regarding sequence-structure relationships[3] and the enormous amount of sequencing information that can be used to build evolutionary trees at the molecular level.[4]

Interestingly, the AlphaFold predictions highlight the importance of intrinsically disordered proteins/regions (IDPs/IDRs). Conservative estimates indicate that roughly 30% of sequences, 30-residues or longer, drawn at random from the

human proteome, are likely to be IDRs.[3,5] A striking feature of structural annotations of the human proteome provided by AlphaFold is the vast number of low and very low confidence regions that overlap with regions that are predicted to be IDRs. This highlights the importance of conformational heterogeneity, which was advocated for in numerous pioneering studies.[6–8] The AlphaFold annotations provide a much-needed impetus for functional and biophysical studies of IDPs/IDRs.[9–11] Indeed, the field of IDPs/IDRs has been gaining in momentum[9,12–23] and will likely receive considerable attention on the heels of the AlphaFold predictions.

Over the past two decades, there has been growing recognition that the sequence-structure–function paradigm must be generalized to account for the role of conformational heterogeneity and conformational dynamics in protein function.[8–10,20,21,24–28] Under standard conditions used to study and characterize protein structures, IDPs/IDRs are defined by conformational heterogeneity, and the preference for heterogeneity is encoded in their amino acid sequences.[11,29–31] Unlike many intrinsically foldable proteins whose functions tend to be ascribed to a specific three-dimensional structure or non-trivial fluctuations centered around a singular structure,[32–35] IDPs/IDRs are best defined by a heterogeneous ensemble of conformations.[36] The overall number of distinct conformations in the ensemble, the conformational heterogeneity, in terms of features such as the sizes and shapes sampled, and the relative importance of different conformations within the ensemble are defined by a combination of sequence-encoded short- and long-range interactions[37,38] as well as the modulation of these interactions by solution conditions.[39,40] Systematic efforts, spanning the past two decades, have led to prediction engines that allow one to identify regions within a protein sequence that are likely to be IDRs.[5,16,41–44] Further, since many (not all) IDRs can undergo coupled folding and binding into context-specific three-dimensional structures,[26,45] predictions have also focused on identifying the proclivity for forming specific structural motifs within IDRs,[46,47] either in their unbound forms or as parts of complexes. Importantly, prediction engines are also available to identify regions that undergo disorder-to-order, order-to-disorder, and disorder-to-disorder transitions upon binding.[44]

Given the advances enabled by AlphaFold, it is now likely obligatory that biologists and biochemists will look up the structures of their favorite proteins. Many of these proteins, especially those involved in signaling, transcription, and coordinating protein–protein interaction networks, are likely to feature large, disordered regions. A typical annotation, shown in Figure 1, will depict large regions as being orange "unstructured" regions of very low confidence. The key questions are: (1) What does one do with this information? And (2) how should one interpret the "unstructured regions" depicted by the AlphaFold annotation? Here, we attempt to answer these questions. As a prelude to offering our perspective regarding the AlphaFold annotations of IDRs, we emphasize the importance of an in-depth reading of the two key papers[1,2] that describe how AlphaFold works. In addition, experts in the fields of structural bioinformatics and machine learning have provided authoritative and accessible accounts of the inner workings of AlphaFold (https://moalquraishi.wordpress.com/2021/07/25/the-alphafold2-method-paper-a-fount-of-good-ideas/). These analyses and the original publications are crucial for understanding what data go in, how they are used, and how a prediction trajectory evolves for each amino acid sequence. As always, the devil is in the details, and these details help offer an educated and critical perspective that keeps us from ascribing more than there should be ascribed to each prediction. One of the major insights leveraged by AlphaFold[1,2] comes from the use of evolutionary covariations that can be extracted from large-scale multiple sequence alignments.[49–54] These discoveries, which go by various names including direct coupling analysis (DCA),[55] have their origins in early work showing that covariation analysis helps with the identification of functionally relevant sectors in protein structures.[56] This is important to acknowledge, because one of the challenges posed by IDRs stems from hypervariability of these regions across orthologs,[57] making it difficult to uncover evolutionary constraints from alignments alone.[58] Instead, the insights regarding evolutionary constraints come from a synthesis of physical chemistry principles, exemplified in sequence-ensemble relationships,[23,59,60] that determine the functions and interoperability of IDRs.

Here, we highlight some of the annotations of disordered regions and the misconceptions that can arise from interpreting information from static structures, especially those with very low confidence scores from AlphaFold predictions that overlap with IDRs. We use this as a segue to highlight the lessons that have been learned from systematic efforts to obtain quantitative descriptions of sequence-ensemble relationships of IDPs/IDRs. We conclude with an outlook for productive interfaces between advances in machine learning and quantitative work on IDPs/IDRs.

Inferences that one might be tempted to draw from predicted structures

Access to predicted structures and annotations regarding the quality of predictions, insightful as they are, can give rise to misconceptions. Figure 1 shows a sampling of structures predicted by AlphaFold. Of relevance are the regions that are

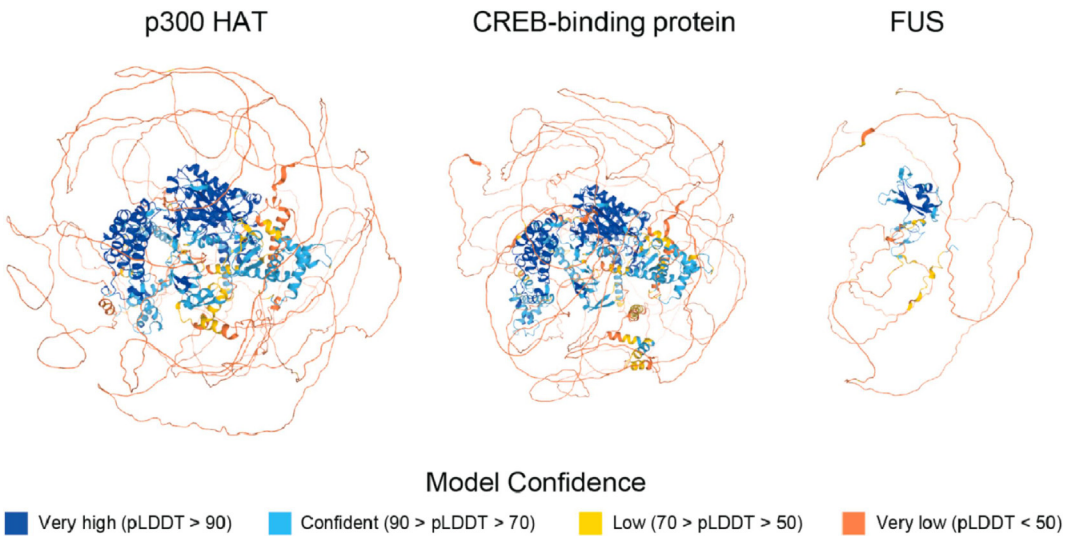


Figure 1. Example showing AlphaFold predicted structures of proteins with long IDRs. p300 HAT (Uniprot: Q09472) and CREB-binding protein (Uniprot: Q92793) are transcriptional co-activators that have hundreds of interaction partners and are involved in many signaling pathways and biological processes.[48] FUS (Uniprot: P35637) is an RNA binding protein involved in many cellular processes including transcriptional activation and RNA processing. Each residue in the sequence is color-coded based on the model confidence score, pLDDT.

assigned very low confidence as quantified by pLDDT.[1,2] This is a per-residue confidence score that is scaled between 0 and 100 and estimates how well the predicted structure would agree with the experimental structure as defined by the predicted Local Distance Difference Test (pLDDT).[1,2] Here, we attempt to answer the following questions: (1) Are all low and very low confidence regions failures of AlphaFold? i.e., do these regions adopt a well-defined three-dimensional structure, but AlphaFold fails in predicting the relevant structures? (2) Do the large clouds/ribbon-like depictions of very low confidence regions in AlphaFold have physical significance in terms of defining the radius of capture of the IDR or the interplay with folded domains?

Are all low and very low confidence regions failures of AlphaFold in that these regions do adopt a well-defined three-dimensional structure that is somehow being missed by AlphaFold? Roughly 30% of the residues across predicted structures in the human proteome tend to have pLDDT scores that are less than 50. This is consistent with estimates regarding the extent of disorder expected within the human proteome. And while it is possible that filtering regions based on pLDDT scores being below 50 can lead to an overestimate of disorder, pLDDT scores have been shown to be a competitive disorder predictor compared to standards in disorder prediction.[5] This implies that a vast majority of the very low confidence regions are likely to be intrinsically disordered, instead of being well-defined, autonomously foldable three-dimensional structures that AlphaFold fails at predicting. Additionally, while some of the low and very low confidence regions might be candidates for coupled folding and binding, whereby a specific structure is acquired in the context of a complex, it is worth noting that the acquired structure might vary with context such as the nature of the binding partner.[24] Further, transitions associated with binding and complex formation can come in different flavors.[44] Therefore, given the prevalence of these regions, the overall implication is that conformational heterogeneity is real and, as has been emphasized over the past two decades, heterogeneity has an important role to play in protein function.

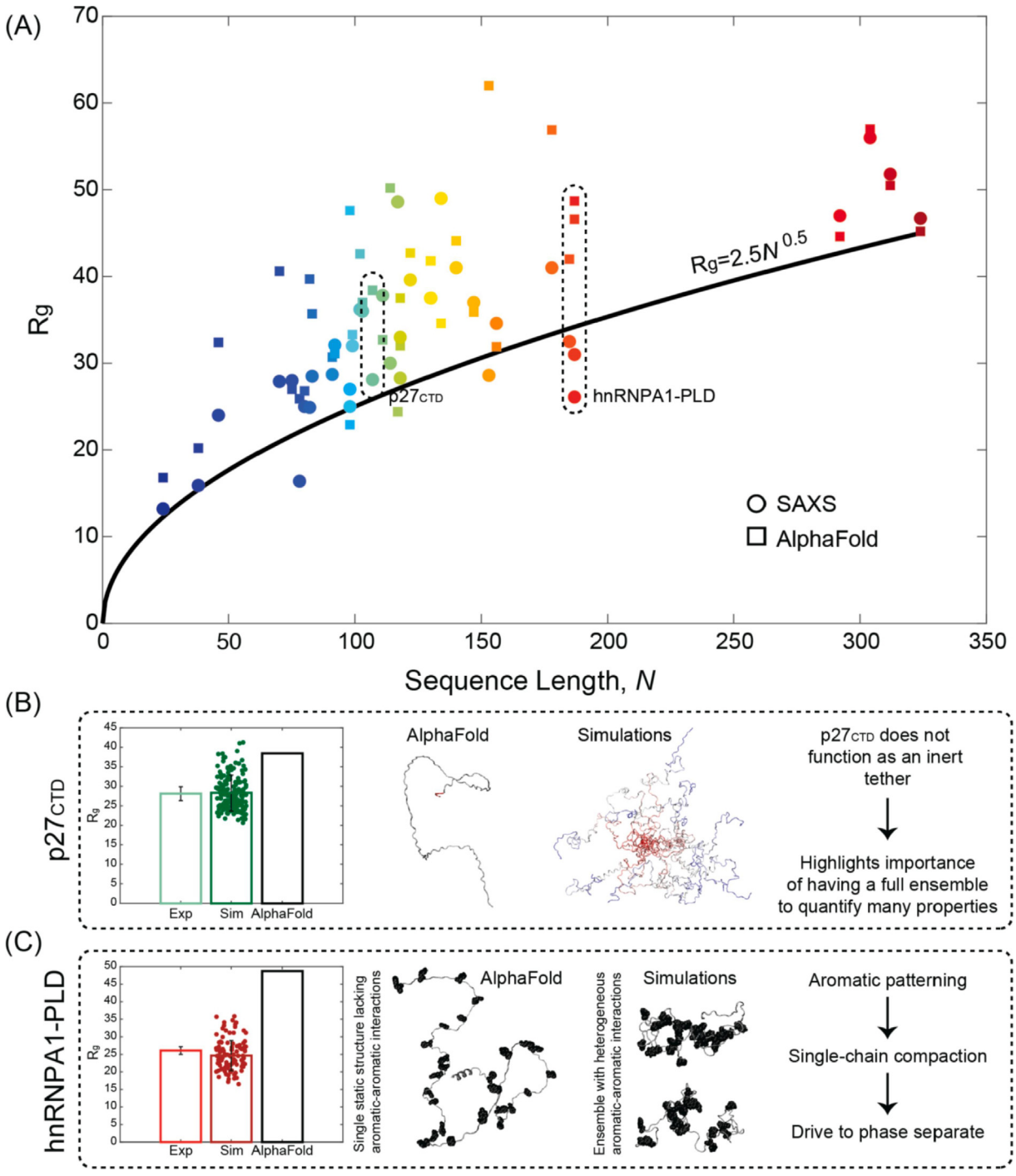
(2) *Do the large clouds/ribbon-like depictions of very low confidence regions in AlphaFold have physical significance in terms of defining the radius of capture of the IDR or the interplay with folded domains?* To answer this question, it is worth emphasizing that intrinsic disorder is not the same as being "unstructured". Instead, disorder implies that a diverse conformational ensemble best describes the region of interest. This ensemble is sequence-specific[30,31] and the extent of heterogeneity[36] as well as the relative preferences of conformations within the ensemble will depend on the primary sequence, solution conditions, and functional contexts. The sequence-specificity of sequence-ensemble relationships cannot be ignored. Accordingly, a single static "structure", even if it is annotated as being a low confidence prediction, cannot be used as a representative conformation that describes the ensemble. Instead, what we need, and is being actively pursued in the IDP field, are quantitative descriptions of conformational ensembles in terms of distribution functions for inter-residue distances and measurements of moments of these distributions.[61,62] Absent such quantitative descriptions,

any static picture showing a "lack of structure" carries no physical meaning other than being a particular choice for annotating regions of conformational heterogeneity. As a result, depictions of the space occupied by the disordered region, or a visual sense of contacts being formed, or inferences regarding motif/residue accessibilities extracted using image processing should not be taken literally because such analyses have zero physical meaning. For example, one must not infer that the envelope one obtains by tracing out the contour of "unstructured" regions has any bearing on the hydrodynamic size[63] or radius of gyration. This is because, the static AlphaFold structure likely overestimates the radius of gyration of IDRs as measured using small angle X-ray scattering (SAXS)[64] (Figure 2(A)). Likewise, the apparent size of the conformational cloud has no bearing on the radius of capture of an IDR, nor does it automatically rule-in or rule out "fly-casting"[65] as the preferred mechanism of molecular recognition through an IDR. The specific choice made by AlphaFold developers for visual depictions of IDRs as being "unstructured" regions should not be taken to mean that this helps us visualize how IDRs might engage in or facilitate interactions with multiple binding partners or enable multivalent interactions. Indeed, it is worth emphasizing that the AlphaFold developers are explicit in making this case, stating that: "In the current dataset, long regions with pLDDT < 50 adopt a readily identifiable ribbon-like appearance, and should not be interpreted as structures but rather as a prediction of disorder."

Although "seeing is believing", the depictions of disordered regions as being "unstructured" clouds do not tell us anything about effective concentrations, preferential localizations within cells, or modes of molecular recognition. All of this requires the type of quantitatively vetted ensembles that often involve multipronged experimental approaches driven by or aided by computations. Efforts to systematize the quantitative characterization of sequence-ensemble relationships are underway, and this has led to the creation of a protein ensemble database (PEDB),[66,67] which went through a major update recently. Prior to highlighting some of the insights we have gleaned from our efforts to quantify sequence-ensemble relationships, it is worth noting that the conformational clouds one observes in AlphaFold predictions may arise from several aspects of the underlying methodology.[1] First, multiple sequence alignments are needed to predict distances between residues. Intrinsically disordered regions often evolve more rapidly than ordered regions and thus alignments of these regions are generally poorer and involve large gaps and extensions of gaps because orthologous IDRs can span vast sequence lengths.[9,23,58,60,68–70] This likely leads to poorly defined distance restraints within the IDRs and between the IDRs and the folded domains. The disparity in the quality of information and restraints that can be gleaned from multiple sequence alignments for intrinsically foldable domains vs. IDRs is a likely contributor to predictions showing sequences with long IDRs forming a "cloud" or envelope around the ordered domains.

Additionally, the cloud-like depictions for structures of IDRs are also likely to represent an artifact of allowing for steric clashes and the steps taken to alleviate these clashes as deemed necessary.[1] Each round of the prediction process involves an energy minimization step using the AMBER99SB force field.[71] These energy minimizations are performed with harmonic restraints applied independently to heavy atoms, to maintain the system near the starting structure. In effect, there are two competing sets of tasks for the energy

Figure 2. Comparisons of inferences based on static structures from AlphaFold predictions to those gleaned from quantitative sequence-ensemble relationships of IDRs. (A) The relationship between sequence length, in terms of number of residues, N, and radius of gyration (R_g) as determined from SAXS experiments (circles) or quantified from the excised single structure from the AlphaFold prediction (squares). The black line denotes the expected R_g for a Flory random coil of the given sequence length. IDRs are extracted from Table 3 from the work of Ruff.[83] In most cases, the R_g values calculated using static structures from AlphaFold deviate from those obtained using experiments. The dashed elliptical envelopes showcase the variation in R_g one observes for sequences of identical lengths. The colors range from cooler to hotter colors as sequence length increases. (B) Comparison of the AlphaFold structure of $p27_{CTD}$ with R_g values measured using SAXS and conformational ensembles from atomistic simulations. Here, each point corresponds to the R_g of a single conformation extracted from the computed conformational ensemble. Atomistic simulations reproduce the average R_g value measured using SAXS. These ensembles show that the conformational ensemble of $p27_{CTD}$ samples a large range of shapes and sizes. (C) Comparison of the AlphaFold "structure" of prion-like low complexity domain from hnRNP-A1 (A1-LCD) with R_g values measured using SAXS and conformational ensembles from atomistic simulations. Highlighted are two representative conformations from atomistic simulations that show heterogeneous interactions between pairs of sequence-proximal or distal aromatic residues (black, space filling). Interactions between aromatic residues are essentially absent in the AlphaFold structure.



minimization process: maintain proximity to the starting structure whilst alleviating steric clashes. The restraints are imposed using multiple harmonic potentials with force constants of 10 kcal/mol-Å^2. Once the minimization has converged, the method then determines which residues still contain violations. The process is iterative, with each step incorporating more restraints, and continues until all steric clashes are alleviated. In the AMBER99SB forcefield, van der Waals interactions are modeled using the Lennard-Jones 12-6 potential. Alleviation of steric clashes leads to a steep decrease in the potential energy, and often the local minimum will be situated closest to the region of the steepest drop in energy.[72,73] Further, given that the residues are treated as an ideal gas with respect to the global frame of reference, the use of centrosymmetric potentials will lead to an isotropic expansion of the

disordered regions that ended up with significant steric clashes during the process used to generate the penultimate structure.

Next, to make our point regarding the importance of accurate, and rigorously vetted conformational ensembles, we showcase a few examples from our work and those of our colleagues regarding insights gained from ensembles for specific IDRs that were generated using atomistic simulations based on the ABSINTH implicit solvation model and forcefield paradigm[74,75] and refined, where necessary, using experiments or reweighted using experimentally derived restraints.[76,77] These illustrations are not intended to highlight our approaches as the chosen ones to pursue. Instead, given that we have ready access to the information we have generated, we use it to showcase the level of quantitative insights one can obtain regarding sequence-ensemble relationships[78] and to highlight the fact that in addition to being sequence-specific, the features of conformational ensembles can only be described in terms of single- or multi-parameter distribution functions and/or moments of these distributions.[79] Accordingly, a static structure either drawn at random from the ensemble or generated using a specific rendering or optimization procedure is neither a qualitative nor quantitative representation of conformational heterogeneity.

Examples of sequence-ensemble descriptions obtained from atomistic simulations vetted by comparisons to multipronged experiments

Sequence-ensemble relationships of most IDRs are obtained by excising the IDRs from their contexts and studying them as autonomous units *i.e.*, as IDPs. The transferability of sequence-ensemble relationships obtained for IDRs characterized as IDPs to IDRs when tethered as tails or linkers in the context of folded domains has been interrogated recently.[80] Although there are clear context dependent modulations of intrinsic sequence-ensemble relationships, these modulations range from being negligible to predictable based on the sequence features of IDRs and the types of folded domains to which IDRs are tethered.[80,81] Therefore, we propose that the growing repository of vetted ensembles for different IDRs, even those studied as IDPs, will likely become a major driver, aided by improved computational methods and advances in machine learning,[82] to put IDRs back into their natural contexts and assess how sequence-ensemble relationships of IDRs are modulated by the context of being tethered to folded domains. A major undertaking in this direction will likely enable improvements in identifying sequence features that determine sequence-ensemble-function relationships of distinct IDRs. With this optimistic prognosis as a prelude, we highlight a few examples of sequence-ensemble relationships that have been derived from systematic efforts that combine atomistic simulations (in our case, based on the ABSINTH implicit solvation model and forcefield paradigm) and low- as well as high-resolution experiments.

$p27^{Kip1}$

A key inhibitor of the cyclin dependent kinases is the protein $p27^{Kip1}$ (or p27). It was one of the earliest IDPs to have been characterized in solution[84] and is a key component that regulates progression through the cell cycle in mammalian systems. The N-terminal portion (residues 1–95) folds upon binding to Cdk2/cyclin A and this leads to kinase inhibition.[85,86] A sizable fraction of the N-terminal region is predicted to adopt a helical structure by AlphaFold, and this is consistent with experimental characterizations based on nuclear magnetic resonance (NMR) spectroscopy.[85] In contrast, the C-terminal domain, residues 96–198, remains disordered upon binding. This region is involved in a phosphorylation/ubiquitination-signaling cascade in which Y88 phosphorylation leads to intra-complex phosphorylation of T187, which in turn activates additional parts of the signaling pathway ultimately leading to full activation of Cdk2/cyclin A. The C-terminal domain of p27, $p27_{CTD}$, generally shows low pLDDT scores (<70) in the AlphaFold prediction. The low, as opposed to very low, confidence score is misleading in this case. The single static structure of $p27_{CTD}$ predicted by AlphaFold suggests that this region is highly expanded, with few intramolecular contacts (Figure 2(B)). However, detailed characterizations based on SAXS, NMR, paramagnetic spin relaxation enhancement measurements, and ion mobility mass spectrometry suggest that the statistical properties of the conformational ensemble of $p27_{CTD}$ are congruent, on average, with those of Flory random coils.[64] Ensembles that have statistical properties of Flory random coils typically maximize conformational entropy, featuring large-scale conformational fluctuations, a counterbalancing of intra-chain attractions by intra-chain repulsions or a counterbalancing of chain-solvent attractions by intra-chain attractions.[79] Detailed insights regarding the interactions that screen one another and the overall shapes, sizes, and local as well as global conformational fluctuations have emerged from atomistic simulations that were then vetted against experimental data. Together, these results suggest that $p27_{CTD}$ samples a large range of sizes and shapes. Modulating the charge patterning changes the mean size of the conformational ensemble monotonically, consistent with SAXS experiments.[64] However, the efficiency of T187 phosphorylation does not change monotonically with changes to R_g. Instead, closer scrutiny shows that the IDR does not behave as an inert tether that influences the radius of capture alone. Instead, there is a cryptic motif that is proxi-

mal to the primary substrate motif that regulates the efficiency of T187 phosphorylation.[64] These results showcase the intricate interplay amongst global dimensions, amplitudes of conformational fluctuations, and local sequence features that regulate the enzymatic efficiency when the substrate is part of an IDR.

The prion-like complexity domain from hnRNPA1

Many proteins with low complexity prion-like domains (PLDs) are drivers of phase separation.[87] The sequence composition of the PLD from hnRNP-A1 (residues 186–372), corresponds to that of an archetypal PLD, and has been designated as the A1-LCD. For most of this region, the AlphaFold prediction has a pLDDT score that is less than 50. The R_g value we compute using the very low confidence static structure excised from the AlphaFold prediction is considerably higher than values obtained from direct measurements based on SAXS (Figure 2(C)). Integrating SAXS and NMR measurements with simulations, suggests that the ensemble of conformations fluctuate among locally compact regions, globally compact conformations, and expanded conformations.[88] These conformational features are driven by the distributed network of contacts formed between pairs of sequence-proximal or distal aromatic residues and influenced or diluted by the *spacer* regions that are interspersed between aromatic residues (*stickers*).[88] In this case, the conformational ensembles provide insights regarding the cohesive interactions that the A1-LCD can engage in, leading to the postulate, tested through theory, computations, and experiments, that aromatic residues function as cohesive motifs known as stickers that drive single chain compaction as well as phase separation[88] – the latter through a network of intermolecular interactions. Importantly, the detailed atomistic descriptions of conformational ensembles enabled the development of a coarse-grained *stickers*-and-*spacers* model for simulations of A1-LCD phase behavior.[88] This work highlights the importance of having accurate knowledge of sequence-specific conformational ensembles for identifying and manipulating the driving forces for phase separation.

Conformational ensembles of the exon 1 encoded region of huntingtin (Httex1)

Aggregation prone molecules are characterized by complex phase behaviors. High-resolution structural studies of purely monomeric species are challenging because these molecules tend to self-associate readily.[89,90] Therefore, structural readouts are often confounded by polydisperse distributions of oligomers and higher-order assemblies. The sequence of Httex1, excised from the protein huntingtin features a poly-Gln (polyQ) stretch that can vary in length from 18 to 25 residues (wild-type) and go well above 35–40 residues (for mutations associated with Huntington's disease). Through advances in chemical biology and single molecule Förster resonance transfer energy (smFRET) measurements, a series of Httex1 constructs, each with different polyQ lengths, were probed to extract intramolecular FRET efficiency histograms.[91] These measurements were performed at concentrations that were in the low picomolar range – a necessity given the polyQ-mediated self-associations that lead to oligomerization and aggregation at higher concentrations. The experimentally derived FRET efficiencies were used in a Bayesian approach to reweight atomic-level descriptions of conformational ensembles[92,93] obtained using large-scale Monte Carlo simulations based on the ABSINTH implicit solvation model and forcefield paradigm. A curious feature of the conformational ensembles was the discovery that Httex1 has a clear preference for forming tadpole-like conformations.[94] The polyQ domain adopts globular conformations, wherein the radius of gyration scales as $n^{1/3}$ where n is the number of Gln residues in the polyQ tract. The Pro-rich C-terminal region forms a semi-flexible tail. Conformational heterogeneity of this ensemble derives from the diversity of compact conformations within the polyQ domain and the distribution of internal distances and orientations that characterize the Pro-rich region. Taken together with complementary efforts focused on characterizing Httex1 phase behavior, the results obtained from multipronged studies of monomeric Httex1 predict that the growth of the "sticky head" with polyQ length mediates homotypic interactions and modulates heterotypic interactions in the cellular context.[94] These changes, as opposed to sharp conformational transitions as a function of polyQ length, are likely to influence the gain-of-function interactions that influence disease progression and neuronal death.

Conclusions

Clearly, the AlphaFold predictions and structural annotations of the human proteome mark an epoch in structural bioinformatics that will fuel several key advances in structural and molecular biology. The predictions validate roughly two decades of intense efforts that have focused on identifying IDRs, highlighting their relevance, and the importance of efforts focused on developing quantitative sequence-ensemble relationships that can be used to discern how conformational heterogeneity and the dynamics of interconversion between distinct conformations contribute to protein function. The successes of the AlphaFold approach will undoubtedly invite considerable attention and investment into supervised learning approaches and the adoption of deeply nested neural networks for quantitative descriptions of sequence-ensemble relationships of IDRs. These

efforts, which will have to go hand in hand with increased throughput and complexities of quantitative studies that combine theory, computations, and an assortment of low, medium, and high-resolution experiments, hold the promise of understanding how IDRs work in synergy with autonomously folded domains to influence protein function, and higher-order interactions. With the appetite for static structural annotations now hopefully satiated, we hope that dedicated efforts will be invested in rigorous statistical physics-based descriptions of conformational heterogeneity combined with an interface with machine learning to uncover how the totality of information encoded in conformational ensembles influences protein function.

CRediT authorship contribution statement

Kiersten M. Ruff: Formal analysis, Writing – original draft, Writing - review & editing. **Rohit V. Pappu:** Conceptualization, Writing – original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to our collaborators (Richard Kriwacki, Danny Hatters, Hilal Lashuel, Edward Lemke, and Tanja Mittag), and past and current members of the Pappu lab for their insights, expertise, and efforts in developing and interpreting quantitative sequence-ensemble relationships for IDPs/IDRs. The current perspective benefited from inputs provided by Alan Chen, Furqan Dar, Alex Holehouse, Matthew King, Min Kyung Shinn, and Andreas Vitalis.

Received 2 August 2021;
Accepted 12 August 2021;
Available online 18 August 2021

Keywords:
AlphaFold;
intrinsically disordered proteins;
cautionary notes

References

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*,.
2. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., et al., (2021). Highly accurate protein structure prediction for the human proteome. *Nature*,.
3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
4. Harms, M.J., Thornton, J.W., (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.*, **20**, 360–366.
5. Necci, M., Piovesan, D., Hoque, M.T., Walsh, I., Iqbal, S., Vendruscolo, M., et al., (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*, **18**, 472–481.
6. Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I., Wright, P.E., (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci.*, **93**, 11504.
7. Radhakrishnan, I., Pérez-Alvarado, G.C., Parker, D., Dyson, H.J., Montminy, M.R., Wright, P.E., (1999). Structural analyses of CREB-CBP transcriptional activator-coactivator complexes by NMR spectroscopy: implications for mapping the boundaries of structural domains11Edited by F. E. Cohen. *J. Mol. Biol.*, **287**, 859–865.
8. Wright, P.E., Dyson, H.J., (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
9. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., et al., (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
10. Milles, S., Jensen, M.R., Lazert, C., Guseva, S., Ivashchenko, S., Communie, G., et al., (2018). An ultraweak interaction in the intrinsically disordered replication machinery is essential for measles virus function. *Sci. Adv.*, **4**, eaat7778.
11. Jensen, M.R., Communie, G., Ribeiro, E.A., Martinez, N., Desfosses, A., Salmon, L., et al., (2011). Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl. Acad. Sci.*, **108**, 9839.
12. Uversky, V.N., (2021). Recent developments in the field of intrinsically disordered proteins: intrinsic disorder-based emergence in cellular biology in light of the physiological and pathological liquid-liquid phase transitions. *Annu. Rev. Biophys.*, **50**, 135–156.
13. Sadar, M.D., (2020). Discovery of drugs that directly target the intrinsically disordered region of the androgen receptor. *Expert Opin. Drug Discov.*, **15**, 551–560.
14. Rehman, A.U., Rahman, M.U., Arshad, T., Chen, H.-F., (2019). Allosteric modulation of intrinsically disordered proteins. In: Zhang, J., Nussinov, R. (Eds.), *Protein Allostery in Drug Discovery*,. Singapore, Springer Singapore, pp. 335–357.
15. Gianni, S., Jemth, P., (2019). Affinity versus specificity in coupled binding and folding reactions. *Protein Eng. Des. Sel.*, **32**, 355–357.
16. Liu, Y., Wang, X., Liu, B., (2017). A comprehensive review and comparison of existing computational methods for

intrinsically disordered protein and region prediction. *Briefings Bioinf.*, **20**, 330–346.
17. Stephens, A.D., Zacharopoulou, M., Kaminski Schierle, G. S., (2019). The cellular environment affects monomeric α-Synuclein Structure. *Trends Biochem. Sci.*, **44**, 453–466.
18. Levengood, J.D., Tolbert, B.S., (2019). Idiosyncrasies of hnRNP A1-RNA recognition: Can binding mode influence function. *Semin. Cell Dev. Biol.*, **86**, 150–161.
19. Roschger, C., Cabrele, C., (2017). The Id-protein family in developmental and cancer-associated pathways. *Cell Commun. Signal.*, **15**, 7.
20. Dyson, H.J., Wright, P.E., (2016). Role of intrinsic protein disorder in the function and interactions of the transcriptional coactivators CREB-binding protein (CBP) and p300*. *J. Biol. Chem.*, **291**, 6714–6722.
21. Guharoy, M., Bhowmick, P., Tompa, P., (2016). Design principles involving protein disorder facilitate specific substrate selection and degradation by the ubiquitin-proteasome system*. *J. Biol. Chem.*, **291**, 6723–6731.
22. Ying, J., Roche, J., Bax, A., (2014). Homonuclear decoupling for enhancing resolution and sensitivity in NOE and RDC measurements of peptides and proteins. *J. Magn. Reson.*, **241**, 97–102.
23. Zarin, T., Strome, B., Nguyen Ba, A.N., Alberti, S., Forman-Kay, J.D., Moses, A.M., (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife*, **8**, e46883
24. Wright, P.E., Dyson, H.J., (2009). Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.
25. Varadi, M., Zsolyomi, F., Guharoy, M., Tompa, P., (2015). Functional advantages of conserved intrinsic disorder in RNA-binding proteins. *PLoS ONE*, **10**, e0139731
26. Miskei, M., Antal, C., Fuxreiter, M., (2016). FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.*, **45**, D228–D235.
27. Tóth-Petróczy, Á., Oldfield, C.J., Simon, I., Takagi, Y., Dunker, A.K., Uversky, V.N., et al., (2008). Malleable machines in transcription regulation: The mediator complex. *PLoS Comput. Biol.*, **4**, e1000243
28. Schneider, R., Maurin, D., Communie, G., Kragelj, J., Hansen, D.F., Ruigrok, R.W.H., et al., (2015). Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. *J. Am. Chem. Soc.*, **137**, 1220–1229.
29. Sterckx, Y.G.J., Volkov, A.N., Vranken, W.F., Kragelj, J., Jensen, M.R., Buts, L., et al., (2014). Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, **22**, 854–865.
30. Das, R.K., Ruff, K.M., Pappu, R.V., (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **32**, 102–112.
31. Holehouse, A.S., Das, R.K., Ahad, J.N., Richardson, M.O., Pappu, R.V., (2017). CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.*, **112**, 16–21.
32. Jain, R.K., Ranganathan, R., (2004). Local complexity of amino acid interactions in a protein core. *Proc. Natl. Acad. Sci.*, **101**, 111.
33. Hekstra, D.R., White, K.I., Socolich, M.A., Henning, R.W., Šrajer, V., Ranganathan, R., (2016). Electric-field-stimulated protein mechanics. *Nature*, **540**, 400–405.
34. Bowman, G.R., Geissler, P.L., (2012). Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci.*, **109**, 11681.
35. Zimmerman, M.I., Porter, J.R., Ward, M.D., Singh, S., Vithani, N., Meller, A., et al., (2021). SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chem.*, **13**, 651–659.
36. Lyle, N., Das, R.K., Pappu, R.V., (2013). A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.*, **139**, 09B607_601.
37. Parigi, G., Rezaei-Ghaleh, N., Giachetti, A., Becker, S., Fernandez, C., Blackledge, M., et al., (2014). Long-range correlated dynamics in intrinsically disordered proteins. *J. Am. Chem. Soc.*, **136**, 16201–16209.
38. Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M.R., Zweckstetter, M., et al., (2010). NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.*, **132**, 8407–8418.
39. Holehouse, A.S., Sukenik, S., (2020). Controlling structural bias in intrinsically disordered proteins using solution space scanning. *J. Chem. Theory Comput.*, **16**, 1794–1805.
40. Moses, D., Yu, F., Ginell, G.M., Shamoon, N.M., Koenig, P. S., Holehouse, A.S., et al., (2020). Revealing the hidden sensitivity of intrinsically disordered proteins to their chemical environment. *J. Phys. Chem. Lett.*, **11**, 10131–10136.
41. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., et al., (2020). MobiDB: Intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.
42. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T., (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
43. Peng, Z., Mizianty, M.J., Kurgan, L., (2014). Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins: Struct. Funct. Bioinf.*, **82**, 145–158.
44. Miskei, M., Horvath, A., Vendruscolo, M., Fuxreiter, M., (2020). Sequence-based prediction of fuzzy protein interactions. *J. Mol. Biol.*, **432**, 2289–2303.
45. Fuxreiter, M., Tompa, P., (2012). Fuzzy complexes: A more stochastic view of protein function. In: Fuxreiter, M., Tompa, P. (Eds.), *Fuzziness: Structural Disorder in Protein Complexes*,. Springer US, New York, NY, pp. 1–14.
46. Mészáros, B., Simon, I., Dosztányi, Z., (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376
47. Malhis, N., Gsponer, J., (2015). Computational identification of MoRFs in protein sequences. *Bioinformatics*, **31**, 1738–1744.
48. Bedford, D.C., Kasper, L.H., Fukuyama, T., Brindle, P.K., (2010). Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases. *Epigenetics*, **5**, 9–15.
49. Shindyalov, I.N., Kolchanov, N.A., Sander, C., (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng. Des. Sel.*, **7**, 349–358.

50. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., et al., (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, **6**, e28766
51. Marks, D.S., Hopf, T.A., Sander, C., (2012). Protein structure prediction from sequence variation. *Nature Biotechnol.*, **30**, 1072–1080.
52. Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., et al., (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, **3**, e03430.
53. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., et al., (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnol.*, **35**, 128–135.
54. de Juan, D., Pazos, F., Valencia, A., (2013). Emerging methods in protein co-evolution. *Nature Rev. Genet.*, **14**, 249–261.
55. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., et al., (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108**, E1293.
56. Lockless, S.W., Ranganathan, R., (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295.
57. Cohan, M.C., Pappu, R.V., (2020). Making the case for disordered proteins and biomolecular condensates in bacteria. *Trends Biochem. Sci.*, **45**, 668–680.
58. Ba, A.N.N., Yeh, B.J., Van Dyk, D., Davidson, A.R., Andrews, B.J., Weiss, E.L., et al., (2012). Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal.*, **5** rs1–rs1.
59. Cohan, M.C., Ruff, K.M., Pappu, R.V., (2019). Information theoretic measures for quantifying sequence-ensemble relationships of intrinsically disordered proteins. *Protein Eng. Des. Sel.*, **32**, 191–202.
60. Zarin, T., Tsai, C.N., Nguyen Ba, A.N., Moses, A.M., (2017). Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci.*, **114**, E1450–E1459.
61. Mao, A.H., Crick, S.L., Vitalis, A., Chicoine, C.L., Pappu, R. V., (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.*, **107**, 8183–8188.
62. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., Schuler, B., (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci.*, **109**, 16155.
63. Sherry, K.P., Das, R.K., Pappu, R.V., Barrick, D., (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci.*, **114**, E9243–E9252.
64. Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., Pappu, R.V., (2016). Cryptic sequence features within the disordered protein $p27^{Kip1}$ regulate cell cycle signaling. *Proc. Natl. Acad. Sci.*,, 201516277.
65. Shoemaker, B.A., Portman, J.J., Wolynes, P.G., (2000). Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proc. Natl. Acad. Sci.*, **97**, 8868.
66. Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., et al., (2013). pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, **42**, D326–D335.
67. Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., et al., (2020). PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, **49**, D404–D411.
68. Brown, C.J., Johnson, A.K., Dunker, A.K., Daughdrill, G. W., (2011). Evolution and disorder. *Curr. Opin. Struct. Biol.*, **21**, 441–446.
69. Bertagna, A., Toptygin, D., Brand, L., Barrick, D., (2008). The effects of conformational heterogeneity on the binding of the Notch intracellular domain to effector proteins: a case of biologically tuned disorder. *Biochem. Soc. Trans.*, **36**, 157–166.
70. Moesa, H.A., Wakabayashi, S., Nakai, K., Patil, A., (2012). Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol. BioSyst.*, **8**, 3262–3273.
71. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C., (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Funct. Bioinf.*, **65**, 712–725.
72. Pappu, R.V., Rose, G.D., (2002). A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.*, **11**, 2437–2455.
73. Stillinger, F.H., Weber, T.A., (1985). Inherent structure theory of liquids in the hard-sphere limit. *J. Chem. Phys.*, **83**, 4767–4775.
74. Vitalis, A., Pappu, R.V., (2009). ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.*, **30**, 673–699.
75. Radhakrishnan, A., Vitalis, A., Mao, A.H., Steffen, A.T., Pappu, R.V., (2012). Improved atomistic Monte Carlo simulations demonstrate that poly-l-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. *J. Phys. Chem. B*, **116**, 6862–6871.
76. Choi, J.M., Pappu, R.V., (2019). Improvements to the ABSINTH force field for proteins based on experimentally derived amino acid specific backbone conformational statistics. *J. Chem. Theory Comput.*, **15**, 1367–1382.
77. Choi, J.M., Pappu, R.V., (2019). Experimentally derived and computationally optimized backbone conformational statistics for blocked amino acids. *J. Chem. Theory Comput.*, **15**, 1355–1366.
78. Shea, J.E., Best, R.B., Mittal, J., (2021). Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **67**, 219–225.
79. Mao, A.H., Lyle, N., Pappu, R.V., (2012). Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J.*, **449**, 307–318.
80. Mittal, A., Holehouse, A.S., Cohan, M.C., Pappu, R.V., (2018). Sequence-to-conformation relationships of disordered regions tethered to folded domains of proteins. *J. Mol. Biol.*, **430**, 2403–2421.
81. Bugge, K., Brakti, I., Fernandes, C.B., Dreier, J.E., Lundsgaard, J.E., Olsen, J.G., et al., (2020). Interactions by disorder – A matter of context. *Front. Mol. Biosci.*, **7**
82. R.J. Emenecker, D. Griffith, A.S. Holehouse, metapredict: a fast, accurate, and easy-to-use cross-platform predictor

of consensus disorder, bioRxiv, 2021.2005.2030.446349 (2021).
83. Ruff, K.M., (2020). Predicting conformational properties of intrinsically disordered proteins from sequence. In: Kragelund, B.B., Skriver, K. (Eds.), *Intrinsically Disordered Proteins: Methods and Protocols*,. Springer US, New York, NY, pp. 347–389.
84. Lacy, E.R., Filippov, I., Lewis, W.S., Otieno, S., Xiao, L., Weiss, S., et al., (2004). p27 binds cyclin–CDK complexes through a sequential mechanism involving binding-induced protein folding. *Nature Struct. Mol. Biol.*, **11**, 358–364.
85. Sivakolundu, S.G., Bashford, D., Kriwacki, R.W., (2005). Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/Cyclin A-bound conformation. *J. Mol. Biol.*, **353**, 1118–1128.
86. Lacy, E.R., Wang, Y., Post, J., Nourse, A., Webb, W., Mapelli, M., et al., (2005). Molecular basis for the specificity of p27 toward Cyclin-dependent kinases that regulate cell division. *J. Mol. Biol.*, **349**, 764–773.
87. Martin, E.W., Mittag, T., (2018). Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry*, **57**, 2478–2487.
88. Martin, E.W., Holehouse, A.S., Peran, I., Farag, M., Incicco, J.J., Bremer, A., et al., (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, **367**, 694–699.
89. Morató, A., Elena-Real, C.A., Popovic, M., Fournet, A., Zhang, K., Allemand, F., et al., (2020). Robust cell-free expression of sub-pathological and pathological huntingtin Exon-1 for NMR studies. General approaches for the isotopic labeling of low-complexity proteins. *Biomolecules*, **10**, 1458.
90. Urbanek, A., Popovic, M., Morató, A., Estaña, A., Elena-Real, C.A., Mier, P., et al., (2020). Flanking regions determine the structure of the poly-glutamine in Huntingtin through mechanisms common among glutamine-rich human proteins. *Structure*, **28**, 733–746. e735.
91. Warner, J.B., Ruff, K.M., Tan, P.S., Lemke, E.A., Pappu, R. V., Lashuel, H.A., (2017). Monomeric Huntingtin Exon 1 has similar overall structural features for wild-type and pathological polyglutamine lengths. *J. Am. Chem. Soc.*, **139**, 14456–14469.
92. Köfinger, J., Stelzl, L.S., Reuter, K., Allande, C., Reichel, K., Hummer, G., (2019). Efficient ensemble refinement by reweighting. *J. Chem. Theory Comput.*, **15**, 3390–3401.
93. Leung, H.T.A., Bignucolo, O., Aregger, R., Dames, S.A., Mazur, A., Bernèche, S., et al., (2016). A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *J. Chem. Theory Comput.*, **12**, 383–394.
94. Newcombe, E.A., Ruff, K.M., Sethi, A., Ormsby, A.R., Ramdzan, Y.M., Fox, A., et al., (2018). Tadpole-like conformations of Huntingtin Exon 1 are characterized by conformational heterogeneity that persists regardless of polyglutamine length. *J. Mol. Biol.*, **430**, 1442–1458.