AlphaFold2: A role for disordered protein prediction?

Carter J. Wilson,[†,‡] Wing-Yiu Choy,[*,¶] and Mikko Karttunen[*,§,‖,‡]

†*Department of Mathematics, The University of Western Ontario, 1151 Richmond Street, Canada, N6A 5B7*

‡*Centre for Advanced Materials and Biomaterials Research, The University of Western Ontario, 1151 Richmond Street, London, Ontario, Canada, N6A 5B7*

¶*Department of Biochemistry, The University of Western Ontario, 1151 Richmond Street, Canada, N6A 5C1*

§*Department of Chemistry, The University of Western Ontario, 1151 Richmond Street, Canada, N6A 3K7*

‖*Department of Physics and Astronomy, The University of Western Ontario, 1151 Richmond Street, Canada, N6A 5B7*

E-mail: jchoy4@uwo.ca; mkarttu@uwo.ca

Abstract

The development of AlphaFold2 was a paradigm-shift in the structural biology community; herein we assess the ability of AlphaFold2 to predict disordered regions against traditional sequence-based disorder predictors. We find that a näive use of Dictionary of Secondary Structure of Proteins (DSSP) to separate ordered from disordered regions leads to a dramatic overestimation in disorder content, and that the predicted Local Distance Difference Test (pLDDT) provides a much more rigorous metric. In addition, we show that even when used for disorder prediction, conventional predictors can

outperform the pLDDT in disorder identification, and note an interesting relationship between the pLDDT and secondary structure, that may explain our observations, and hints at a broader application of the pLDDT to IDP dynamics.

Introduction

Predicting the three dimensional structure of a protein from its primary amino acid sequence is a grand challenge in molecular structural biology dating back to the late 1950's[1,2]. About a year ago in late autumn 2020, AlphaFold2, a deep-learning program, provided a a paradigm-shift in this problem[3]. Not only did it outperform all other groups at the 14th Critical Assessment of protein Structure Prediction (CASP14)[3], but it did so with an astonishing accuracy and a large margin, and consequently caused immediate enthusiasm in related fields such as drug development[4].

The full problem of protein folding is however, multi-faceted, and despite AlphaFold's stellar success, many problems and open questions remain. As has already been pointed out by several authors[5–7], dynamics of protein folding remains a formidable problem; prediction of the folding pathways, effects of mutations, the solution environment, aggregation and, as a very particular category, intrinsically disordered proteins (IDPs).

IDPs remain a major challenge since they are almost entirely devoid of native structure and also because they function primarily as a conformational ensemble[8–11] with folding free energy landscapes that are relatively flat[12–14]. This is a direct consequence of their amino acid sequences[15–17], in particular the enrichment of disorder-promoting residues over and above order-promoting ones[18–21]. The application of AlphaFold2 to the prediction of disordered regions and proteins has only briefly been discussed in the literature[6,7,22], and its performance against traditional predictor methods is currently absent.

In light of the recent publication of the Critical Assessment of protein Intrinsic Disorder (CAID) benchmark[23], detailing the performance of over three dozen sequence

based disorder predictors and their datasets, we saw an excellent opportunity to benchmark AlphaFold2. Herein we compare the performance of AlphaFold2 to the top performing sequence predictors as determined at CAID. We find that a näive application of structure assignment provided by DSSP [24], the primary method for assigning secondary structure based on protein, geometry for the determination of disordered regions, is inaccurate.

The predicted Local Distance Difference Test (pLDDT), which is correlated to the confidence of the structure prediction, provides a better metric for identifying ordered and disordered regions. Furthermore, we find that traditional predictors are capable of outperforming AlphaFold2 in disorder prediction even when the pLDDT is used. We also show how secondary structure and pLDDT scores are interestingly related, providing a potential explanation for the observed performance discrepancy and suggesting a possible link between IDP dynamics and the pLDDT.

Methodology

Dataset generation

Two datasets were used in this work, DisProt and DisProt-PDB derived from the DisProt database [25]. Both reference sets are based on the CAID benchmark dataset and are composed of 475 targets, annotated between June 2018 and November 2018 (DisProt release 2018_11). Note that this is less than the 646 targets used at CAID because AlphaFold2 predicted structures do not exist for some sequences. In the DisProt reference set, all residues not labeled as disordered (1) are labelled as ordered (0). In the DisProt-PDB set, residues for which structural data are available are labelled ordered, however a disorder assignment in the DisProt database overrides this order assignment. All residues not covered by either DisProt annotation or PDB structures are masked and were excluded from analysis. As a result the DisProt-PDB dataset contains no 'uncertain' residues, all residues considered in this

set have either a DisProt annotation or belong to a PDB structure. Additional details pertaining to dataset construction are provided in Supplementary Information and the full list of proteins, structures and combined disorder data are available at `https://github.com/SoftSimu/AlphaFoldDisorderData`.

AlphaFold2 structures were downloaded from the EMBL database (`https://alphafold.ebi.ac.uk/`) and run using DSSP[24] to assign secondary structure. We assume residues belonging to helices, strands, or H-bond stabilized turns are ordered (0) and all other residues are disordered (1). We refer to this as the DSSP predictor or DSSPp for short.

We also collected the predicted Local Distance Difference Test (pLDDT) for each structure. Every residue in an AlphaFold2 structure is assigned a value, scaled between 0 and 100, that estimates how well the experimental and predicted structure would agree based on the Local Distance Difference Test (lDDT)[3,22,26]. We transform this value according to the equation,

$$\text{tpLDDT} = 1 - \text{pLDDT}/100, \tag{1}$$

as suggested by Tunyasuvunakool *et al.*[22], giving us a pLDDT-based predictor of disorder, where 1 is disordered and 0 is ordered. We refer to this prediction method as the transformed pLDDT or tpLD for short.

We can discretize this pLDDT predictor by classifying a residue with a pLDDT score $\geq n$ as ordered (0) and disordered (1) otherwise; we use pLDDT_n (or pLD_n for short), to indicate this binary predictor. Thresholds for n were chosen based on the Matthews correlation coefficent (MCC), that has been documented to be an excellent metric for assessing the accuracy of binary classifiers[27] and was the approach used at CAID[23].

The CAID dataset contains predictions made by three dozen predictors; we selected the top 10 performing on the DisProt and DisProt-PDB giving a combined non-redundant set of 11 (fIDPnn[28], SPOT-Disorder2[29], RawMSA[30], fIDPln[28], Predisorder[31], AUCpreD[32], SPOT-Disorder1[33], SPOT-Disorder-Single (SPOT-Disorder-S)[34]

, DisoMine[35], AUCpreD-np[32] and ESpritz-D[36]). The sequence predictors provide a score between 0 and 1 inclusive as well as a binary disorder/order assignment. No modification to the classification thresholds for these predictors was attempted. Descriptions of disorder prediction methods are provided in the Supplementary Information of the original CAID paper[23].

For two vectors v and w we compute the RMSD as

$$\mathrm{RMSD} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}|v_i - w_i|^2},$$

where m is the number of elements (residues) in each vector (protein), v and w. Given binary vectors a random predictor has an RMSD of ∼0.7 on a uniform dataset. Receiver operating characteristic (ROC), area under the curve (AUC), precision-recall, F_1-score and correlation analysis were all performed using scikit-learn[37] and kernel density estimates (KDE) analysis was performed in seaborn[38]. Descriptions of statistical methods are provided in Supplementary Information.

Results

pLDDT performs better than näive use of DSSP for disorder prediction

Improved performance with tpLD (Eq. 1) over and against DSSPp is evidenced by the ROC curves and AUC values (Figs. 1a, S1a), as well as the precision-recall (PR) curves and F_{max} values (Figs. 1b, S1b) on both the DisProt-PDB and DisProt datasets (Tables S1 and S2). Thresholds for the binary pLD_n predictor were selected based on the Matthews correlation coefficients which gave values of 76 and 68 for the DisProt and DisProt-PDB datasets respectively (Tables S3 and S4). We refer to these discrete predictors as pLD_{76} and pLD_{68}. Unsurprisingly, these values agree with the minimum distance from the ROC curve to the top left of the plot (i.e. (0,1)) (Fig. 1). The dif-

ference between these two values undoubtedly stems from the nature of the underlying datasets, while DisProt-PDB contains no uncertain residues, Disprot does. For analysis purposes, we opt to use a combined pLDDT metric, denoted pLD_{72} that is the mean of these two. Data using multiple pLDDT values is provided in Tables S1 and S2.

RMSD calculations comparing DSSPp and pLD_{72} demonstrate improved performance for all protein classes, including highly disordered (i.e. $> 95\%$) and highly ordered (i.e. $< 10\%$), irrespective of dataset (Fig. 2). We note that overall RMSD values are markedly lower for the DisProt-PDB dataset, again likely a result of it lacking "uncertain" residues – residues for which no PDB or experimental data exists. Shifts towards lower RMSD irrespective of dataset, or protein length and disorder content, are also evident for pLD_{72} (Figs. S2, S3). Regression analysis revealed stronger correlations between pLD_{72} and the traditional disorder predictors with respect to residue-wise disorder RMSD when compared with DSSPp (Figs. S4–S7).

Considering global disorder content prediction, we find that on the DisProt dataset, pLD_{72} shows slightly better performance than DSSPp, with a lower mean and a more accurate distribution; however, we note that both methods significantly overestimate disorder content (Fig. 3). On the DisProt-PDB dataset, closer agreement between pLD_{72} and DSSPp is evident based on the mean with both methods returning values similar to experiment. The two distributions are, however, notably different. While that produced by pLD_{72} has a peak around 0.15 in close agreement with experiment, the peak in the distribution produced by DSSPp is larger and shifted to a higher value around 0.3. This is all to say that a näive application of DSSP for the prediction of disordered and ordered regions for AlphaFold2 structures, specifically the assumption that helical and strand regions are ordered and coiled regions are unstructured, leads to poorer prediction (i.e., higher RMSD, lower AUC and higher F_{max}) of disordered regions and an overestimation in disorder content.

Sequence predictors can still outperform AlphaFold2 on disorder prediction

Comparing the pLDDT-based and DSSPp predictors to various sequence-based predictors revealed performance differences amongst the methods. Notably, tpLD (Eq. 1) performed exceptionally well on the DisProt-PDB dataset posting the largest F_{max} (0.784) and one of the largest AUC (0.905) values of the methods considered (Fig. 1, Tables S1 and S3). This was also evidenced by pLD_{72} which had the highest MCC (0.701) (Table. S1) and one of the lowest RMSD values (Fig. 2) on the Disprot-PDB dataset. Interestingly, on the DisProt dataset, both tpLD (Eq. 1) and DSSPp performed significantly worse and were readily outperformed by the other predictor methods, in particular fIDPnn (F_{max}: 0.357 (DSSPp), 0.429 (tpLD), 0.457 (fIDPnn); AUC: 0.635 (DSSPp), 0.731 (tpLD), 0.794 (fIDPnn)), which outperformed all other predictors, as evidenced by the ROC, PR, and RMSD analyses. We note that with respect to MCC, pLD_{72} still performed well on both the DisProt and DisProt-PDB datasets achieving scores of 0.310 and 0.697 respectively (Tables S1, S2). In agreement with the CAID results we found that SPOT-Disorder2, fIDPnn, RawMSA and AUCpreD all performed exceptionally well (Figs. 1 and S1, Tables S3 and S4)[23].

Secondary structure codons (SSC) reveal relationships between the pLDDT and secondary structure

In order to explain the discrepancy between the pLDDT-based and DSSP predictors with respect to local and global disorder prediction, we considered how pLDDT values were assigned to the secondary structures. Kernel density estimates (KDE) of the distribution of pLDDT values sampled over all residues reveal a strong left-skew for all but the coil secondary structure which exhibits a right-skewed bimodal distribution with peaks around 94 and 35 (Fig. 4). Residues assigned to β-strand and β-bridge structures are the most likely to be assigned to large pLDDT values, followed by helical and H-bond stabilized turns. To provide a more detailed picture of the dis-

tributions, we introduce the concept of a secondary structure codon (SSC), a triplet describing the local secondary structure at a given residue. Analysis of the distributions of pLDDT values for each SSC revealed that residues predicted to belong to both the ends (HHC/CHH/HHT/THH) and middle (HHH) of helices can have pLDDT values <50 (Fig. S8), this was not observed for residues belonging to the middle (EEE) and ends of β-strands(EEC/CEE/EET/TEE) (Fig. S9). For highly coiled residues (CCC/CCT/TCC), both high (>80) and low (<50) pLDDT values were observed (Figs. S10 and S11).

Discussion

AlphaFold2 has been a paradigm-shift in structural biology, providing a tentative solution to the protein folding problem that has persisted over half a century[1]. Since the time that problem was posed by Perutz and Kendrew, a new class of proteins has been discovered and IDPs have become the focus of much study[8,9,39,40]. Over the past two decades much effort has been devoted to developing methods for identifying disordered regions given the primary sequence of a protein[23,41–45], herein, we assess the applicability of AlphaFold2 to this problem.

We find, and strongly stress, that simply inferring a residue in an AlphaFold2 structure assigned by DSSP to a helical, strand, or H-bond stabilized turn is ordered, and otherwise is disordered, results in an overestimation of disorder content and a poor prediction of disordered regions. Instead, employing the pLDDT, a measure of the expected position error at a given residue and originally purposed to assess inter-domain accuracy, provides a much more accurate metric for determining global and local disorder content. Using the pLDDT as a disorder predictor metric we observe impressive performance on the DisProt-PDB dataset when compared to conventional disorder predictors (Fig. 1). While poorer performance is observed on the plain DisProt dataset, pLDDT does outperform naïve use of DSSP in both cases.

Secondary structure and global disorder analyses point to a potential root of the

prediction discrepancy between pLDDT and DSSP, simply put, for AlphaFold2, not all secondary structures are created equal. AlphaFold2 will readily assign a coiled geometry and a high pLDDT value to the same residue, and conversely assign low pLDDT values to structured regions (Fig. 4). While a DSSP predictor assumes that coils are disordered and helices are ordered, a pLDDT predictor will account for the fact that a coil may be more ordered and a helix more disordered for certain residues in certain proteins. It is this former case that is likely resulting in the improved performance observed for pLDDT and underscores the importance of the nuance provided by this metric for disordered protein prediction.

Second to the problem of predicting the (dis)orderedness of a region within a protein is predicting the structural dynamics and transitions (i.e. order-to-disorder, disorder-to-order, disorder-to-disorder) that an IDP may undergo[41,46]. In light of the secondary structure analysis, the pLDDT may be just such a means for extracting this information, namely the transientness of secondary structures, their potential for transition upon binding and their functional importance. A helix with a low pLDDT may be more transient, existing frequently in a disordered, unfolded state, than a helix with a high pLDDT and conversely, a coiled region with a high pLDDT, may suggest a disorder-order transition and/or its conserved role in some biophysical interaction. We here reiterate that by their very nature IDPs exhibit a high degree of conformational flexibility, allowing them to interact with multiple binding partners in a variety of ways[47–53]. While it is the case that a single, static, AlphaFold2 structure cannot adequately describe these often large conformational ensembles[8–10], the ability of the program to predict with relatively high accuracy the location of disordered regions is nonetheless impressive, and refinement of the training set to account for more accurate disordered structures could further improve performance. In addition, thorough analysis of the pLDDT score as it relates to disorder-order transitions, as well as the local function and dynamics of IDP motifs, may further enhance the utility of AlphaFold2 to the IDP community.

While experimental NMR[54–63], and high-quality molecular simulations[64–75] are

some of the most accurate methods for determining the (dis)ordered nature and dynamics of proteins, fast and computationally efficient methods play an important role. Unlike conventional predictors however, AlphaFold2 supplies both a pLDDT score, that can provide an accurate prediction of protein disorderedness, in addition to a three-dimensional structure, that when taken in tandem, may also provide insight into the underlying dynamics of disordered protein regions.

Conclusion

In this study, we have assessed the ability of AlphaFold2 to predict disordered protein regions. We benchmark the program on the DisProt-PDB and DisProt datasets developed for CAID, and find it to perform quite well, exceeding the performance of 11 traditional predictors on the DisProt-PDB dataset. Furthermore, we observe that the pLDDT score assigned to each residue by AlphaFold2 provides an impressive metric for assessing disorder, far surpassing a näive application of DSSP. Our analysis also reveals a link between secondary structure and the pLDDT score, suggesting that continued research into this metric may reveal a fundamental connection to the dynamics of disordered proteins.

Acknowledgements

The authors thank SharcNet and Compute Canada for computational resources.

Funding

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding. M.K. also thanks the Canada Research Chairs Program for financial support.

Figures

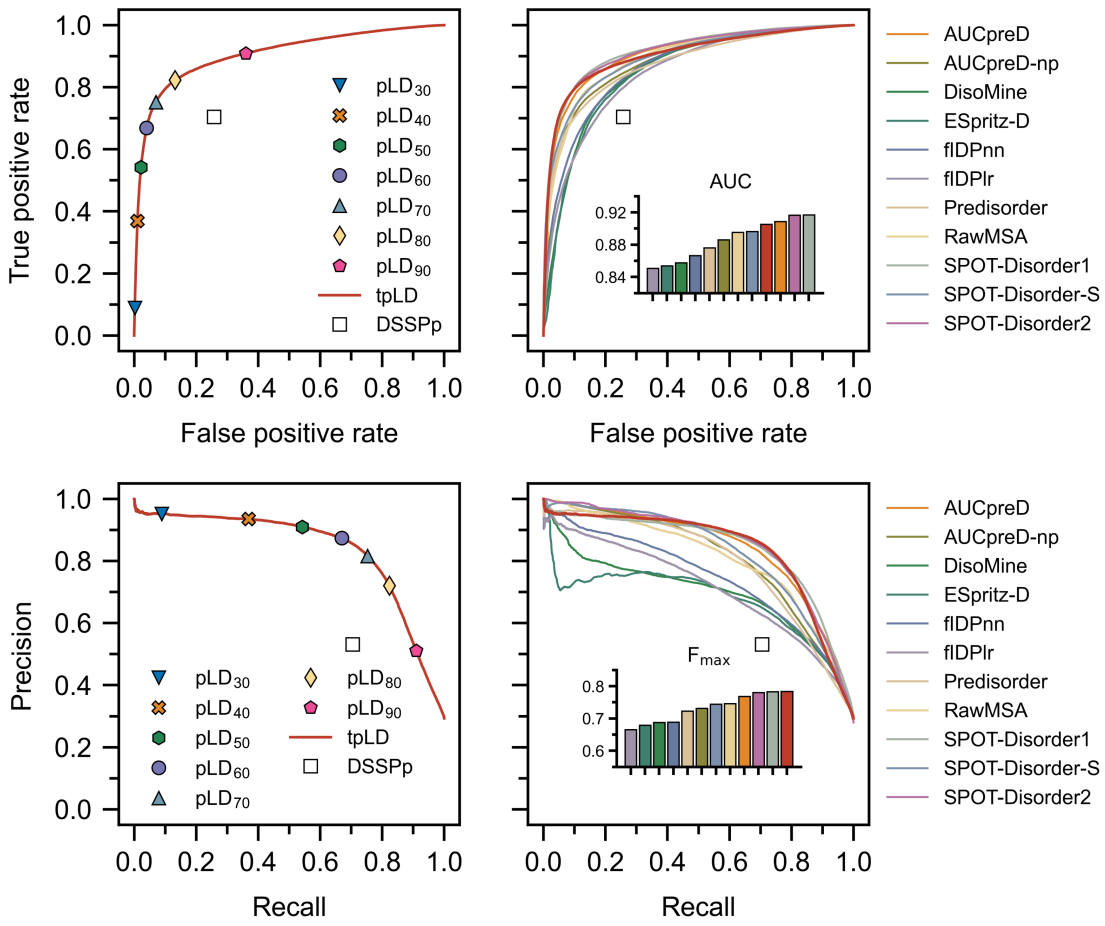

Figure 1: Receiver operating characteristic (ROC) curves (top) and precision-recall (bottom) are depicted for various predictors calculated per-residue on the DisProt-PDB dataset. tpLD (Eq. 1) and various discrete pLD_n predictors are indicated alongside DSSPp. Inset bar plots show the $\mathrm{F}_{\max}$ (top inset) and AUC (bottom inset) for the various predictors on the DisProt-PDB dataset (colors correspond to the legend; red is tpLD). The tpLD predictor resulted in one of the highest AUC values and the highest $\mathrm{F}_{\max}$ on the DisProt-PDB dataset. pLDDT is abbreviated pLD for plotting purposes.

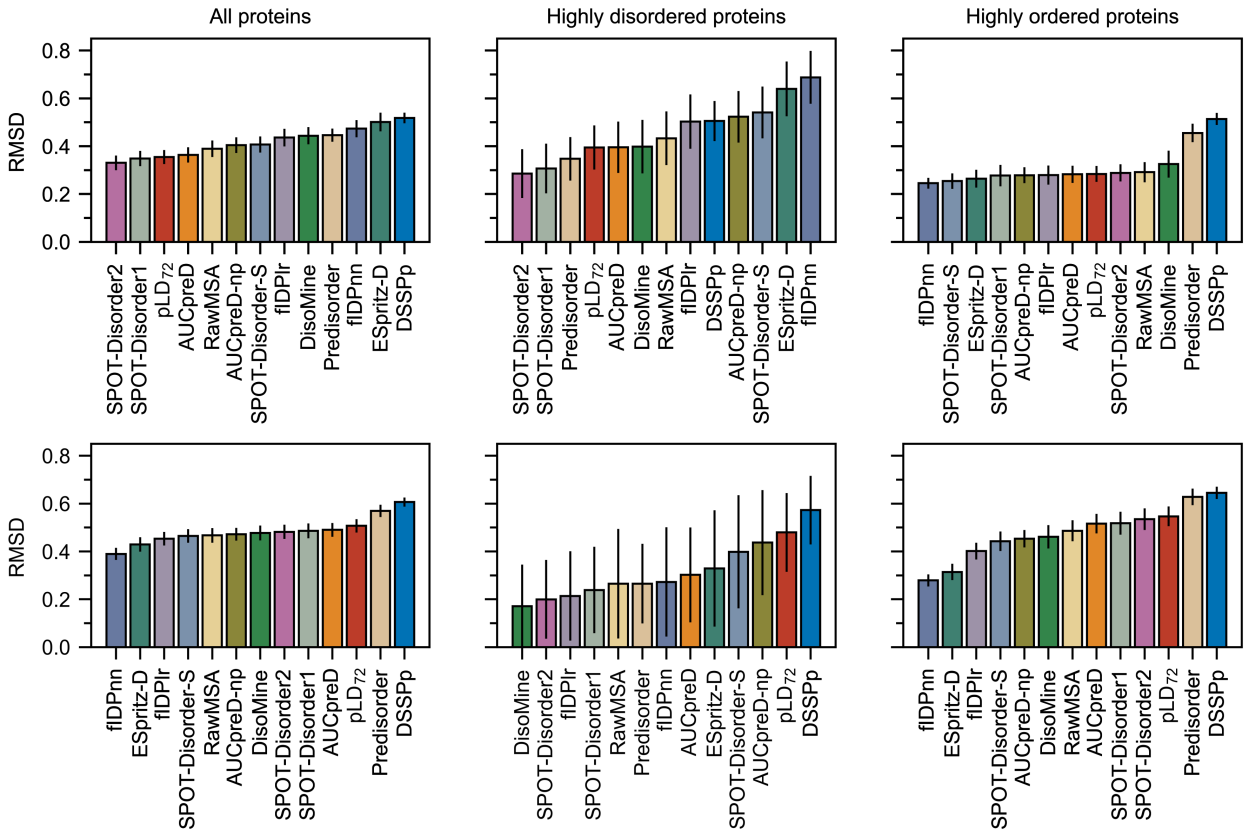

Figure 2: Average RMSD values calculated for the DisProt-PDB (upper) and DisProt (lower) datasets using various prediction methods calculated per-protein. Proteins were assigned to classes (highly disordered i.e. $> 90\%$ disorder and highly ordered i.e. $< 10\%$ disorder) based on datasets; specifically with DisProt-PDB, only residues for which PDB or DisProt data were available are considered in the total disorder calculation. Bootstrapping was used to compute averages and estimate errors with 10,000 samples of size 60. pLD_{72} resulted in lower RMSD values on the DisProt-PDB dataset compared to DSSPp, however both showed much higher RMSD on the DisProt dataset. pLDDT is abbreviated pLD for plotting purposes.

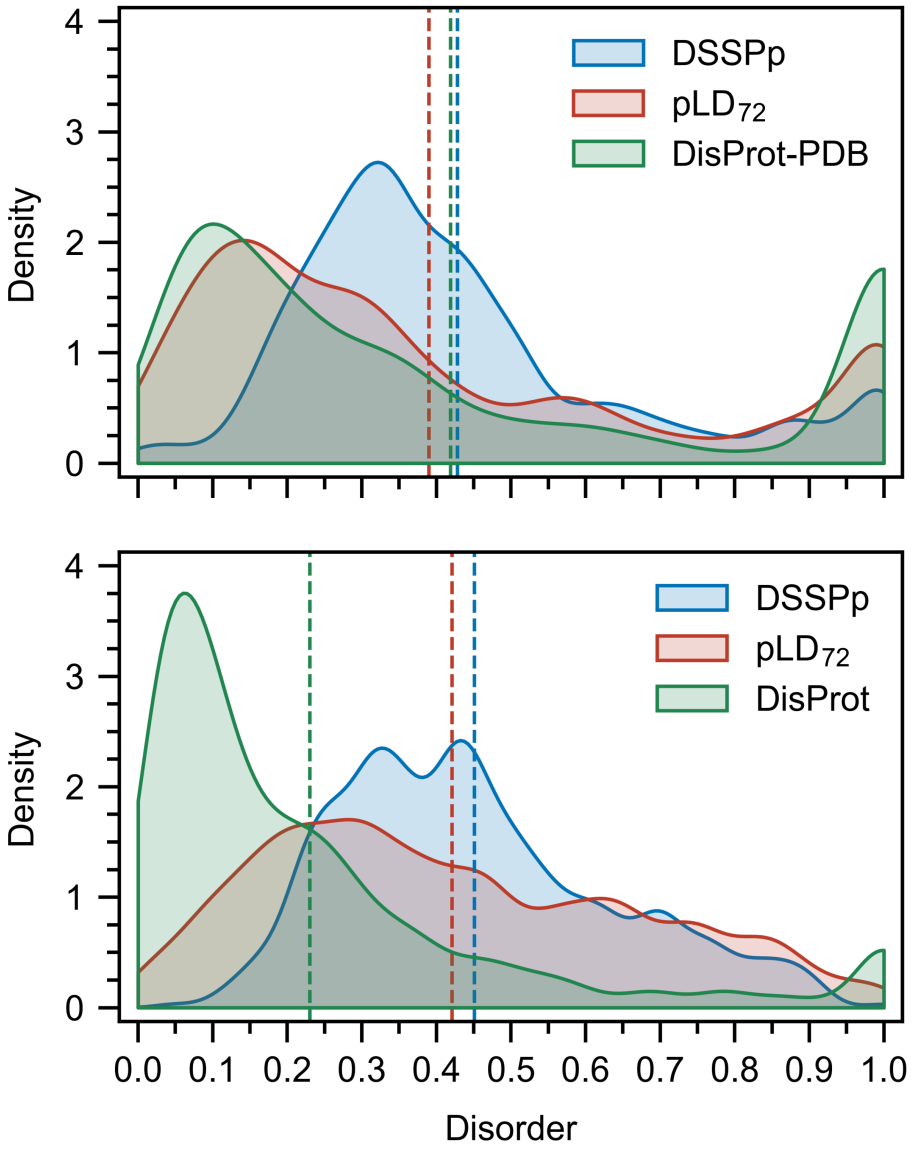

Figure 3: Distribution of disorder content per-protein in the DisProt-PDB and DisProt datasets depicted alongside the distributions predicted by pLD_{72} and DSSPp. Bin-widths were set at 0.5 and bootstrapping was used to compute the distributions and average values (vertical dashed lines) with 10,000 samples of size 60. On the DisProt-PDB dataset close agreement between experiment and pLD_{72} is evident; conversely, on both the DisProt-PDB and DisProt datsets, DSSPp predicted a higher disorder content. pLDDT is abbreviated pLD for plotting purposes.

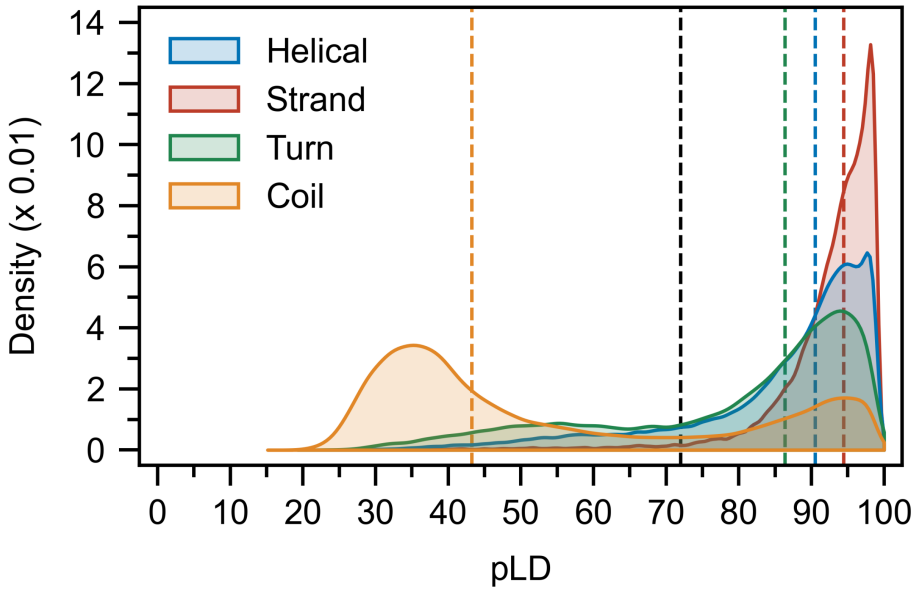

Figure 4: Distribution of pLDDT values per-residue calculated for each secondary structure class. Bin-widths were set at 0.5 and bootstrapping was used to compute the distributions and mean values (colored vertical dashed lines; black dashed line represents pLD_{72}) with 10,000 samples of size 500. A bimodal distribution is evident for the coil structures, and while strand, helical and turn regions are on average assigned to high pLDDT values, residues belonging to each can sample much lower values. pLDDT is abbreviated pLD for plotting purposes.

References

(1) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338*, 1042–1046.

(2) Nassar, R.; Dignon, G. L.; Razban, R. M.; Dill, K. A. The Protein Folding Problem: The Role of Theory. *J. Mol. Biol.* **2021**, 167126.

(3) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(4) Mullard, A. What does AlphaFold mean for drug discovery? *Nat. Rev. Drug Discov.* **2021**,

(5) Serpell, L. C.; Radford, S. E.; Otzen, D. E. AlphaFold: A Special Issue and A Special Time for Protein Science. *J. Mol. Biol.* **2021**, 167231.

(6) Strodel, B. Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, 167182.

(7) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, 167208.

(8) Uversky, V. N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front. Phys.* **2019**, *7*, 10.

(9) DeForte, S.; Uversky, V. N. Intrinsically Disordered Proteins in PubMed: what can the tip of the iceberg tell us about what lies below? *RSC Adv.* **2016**, *6*, 11513–11521.

(10) Lyle, N.; Das, R. K.; Pappu, R. V. A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.* **2013**, *139*, 121907.

(11) Choi, U. B.; Sanabria, H.; Smirnova, T.; Bowen, M. E.; Weninger, K. R. Spontaneous Switching among Conformational Ensembles in Intrinsically Disordered Proteins. *Biomolecules* **2019**, *9*.

(12) Turoverov, K. K.; Kuznetsova, I. M.; Uversky, V. N. The protein kingdom extended: Ordered and Intrinsically Disordered Proteins, their folding, supramolecular complex formation, and aggregation. *Prog. Biophys. Mol. Biol.* **2010**, *102*, 73–84.

(13) Uversky, V. N. Unusual biophysics of Intrinsically Disordered Proteins. *Biochim. Biophys. Acta Proteins Proteom.* **2013**, *1834*, 932–951.

(14) Fisher, C. K.; Stultz, C. M. Constructing ensembles for Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426–431.

(15) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102–112.

(16) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *PNAS* **2013**, *110*, 13392–13397.

(17) Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *PNAS* **2010**, *107*, 8183–8188.

(18) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence complexity of disordered protein. *Proteins Struct. Funct. Bioinf.* **2001**, *42*, 38–48.

(19) Radivojac, P.; Iakoucheva, L. M.; Oldfield, C. J.; Obradovic, Z.; Uversky, V. N.; Dunker, A. K. Intrinsic Disorder and Functional Proteomics. *Biophys. J.* **2007**, *92*, 1439–1456.

(20) Theillet, F.-X.; Kalmar, L.; Tompa, P.; Han, K.-H.; Selenko, P.; Dunker, A. K.; Daughdrill, G. W.; Uversky, V. N. The alphabet of intrinsic disorder. *Intrinsically Disord. Proteins* **2013**, *1*, e24360.

(21) Uversky, V. N. The alphabet of intrinsic disorder. *Intrinsically Disord. Proteins* **2013**, *1*, e24684.

(22) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596.

(23) Necci, M.; ; Piovesan, D.; and, S. C. E. T. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **2021**, *18*, 472–481.

(24) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

(25) Hatos, A.; Hajdu-Soltész, B.; Monzon, A. M.; Palopoli, N.; Álvarez, L.; Aykac-Fas, B.; Bassot, C.; Benítez, G. I.; Bevilacqua, M.; Chasapi, A.; Chemes, L.; Davey, N. E.; Davidović, R.; Dunker, A. K.; Elofsson, A.; Gobeill, J.; Foutel, N. S. G.; Sudha, G.; Guharoy, M.; Horvath, T.; Iglesias, V.; Kajava, A. V.; Kovacs, O. P.; Lamb, J.; Lambrughi, M.; Lazar, T.; Leclercq, J. Y.; Leonardi, E.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Maiani, E.; Manso, J. A.; Marino-

Buslje, C.; Martínez-Pérez, E.; Mészáros, B.; Mičetić, I.; Minervini, G.; Murvai, N.; Necci, M.; Ouzounis, C. A.; Pajkos, M.; Paladin, L.; Pancsa, R.; Papaleo, E.; Parisi, G.; Pasche, E.; Pereira, P. J. B.; Promponas, V. J.; Pujols, J.; Quaglia, F.; Ruch, P.; Salvatore, M.; Schad, E.; Szabo, B.; Szaniszló, T.; Tamana, S.; Tantos, A.; Veljkovic, N.; Ventura, S.; Vranken, W.; Dosztányi, Z.; Tompa, P.; Tosatto, S. C. E.; Piovesan, D. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **2019**, *48*, D269–D276.

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

(27) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*.

(28) Hu, G.; Katuwawala, A.; Wang, K.; Wu, Z.; Ghadermarzi, S.; Gao, J.; Kurgan, L. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **2021**, *12*.

(29) Hanson, J.; Paliwal, K. K.; Litfin, T.; Zhou, Y. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genom. Proteom. Bioinform.* **2019**, *17*, 645–656.

(30) Mirabello, C.; Wallner, B. rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE* **2019**, *14*, e0220182.

(31) Deng, X.; Eickholt, J.; Cheng, J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinform.* **2009**, *10*, 436.

(32) Wang, S.; Ma, J.; Xu, J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* **2016**, *32*, i672–i679.

(33) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2016**, *33*, 685–692.

(34) Hanson, J.; Paliwal, K.; Zhou, Y. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J. Chem. Inf. Model.* **2018**, *58*, 2369–2376.

(35) Orlando, G.; Raimondi, D.; Codice, F.; Tabaro, F.; Vranken, W. Prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics. **2020**,

(36) Walsh, I.; Martin, A. J. M.; Domenico, T. D.; Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **2011**, *28*, 503–509.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(38) Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **2021**, *6*, 3021.

(39) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589–6631.

(40) Uversky, V. N. 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.* **2021**, *50*, 135–156.

(41) Miskei, M.; Horvath, A.; Vendruscolo, M.; Fuxreiter, M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.* **2020**, *432*, 2289–2303.

(42) Peng, Z.; Mizianty, M. J.; Kurgan, L. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* **2013**, *82*, 145–158.

(43) Ward, J.; Sodhi, J.; McGuffin, L.; Buxton, B.; Jones, D. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635–645.

(44) Piovesan, D.; Necci, M.; Escobedo, N.; Monzon, A. M.; Hatos, A.; Mičetić, I.; Quaglia, F.; Paladin, L.; Ramasamy, P.; Dosztányi, Z.; Vranken, W. F.; Davey, N. E.; Parisi, G.; Fuxreiter, M.; Tosatto, S. C. E. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **2020**, *49*, D361–D367.

(45) Liu, Y.; Wang, X.; Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **2017**, *20*, 330–346.

(46) Lindorff-Larsen, K.; Kragelund, B. B. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. 2021.

(47) Freiberger, M. I.; Wolynes, P. G.; Ferreiro, D. U.; Fuxreiter, M. Frustration in Fuzzy Protein Complexes Leads to Interaction Versatility. *J. Phys. Chem. B* **2021**, *125*, 2513–2520.

(48) Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **2014**, *83*, 553–584.

(49) Uversky, V. N. Multitude of binding modes attainable by Intrinsically Disorder Proteins: a portrait gallery of disorder-based complexes. *Chem. Soc. Rev.* **2011**, *40*, 1623–1634.

(50) Sharma, R.; Raduly, Z.; Miskei, M.; Fuxreiter, M. Fuzzy complexes: Specific binding without complete folding. *FEBS Lett.* **2015**, *589*, 2533–2542.

(51) Khan, H.; Cino, E. A.; Brickenden, A.; Fan, J.; Yang, D.; Choy, W.-Y. Fuzzy Complex Formation between the Intrinsically Disordered Prothymosin α and the Kelch Domain of Keap1 Involved in the Oxidative Stress Response. *J. Mol. Biol.* **2013**, *425*, 1011–1027.

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

(53) Arbesú, M.; Iruela, G.; Fuentes, H.; Teixeira, J. a. M. C.; Pons, M. Intramolecular Fuzzy Interactions Involving Intrinsically Disordered Domains. *Front. Mol. Biosci.* **2018**, *5*, 39.

(54) Killoran, R. C.; Sowole, M. A.; Halim, M. A.; Konermann, L.; Choy, W.-Y. Conformational characterization of the intrinsically disordered protein Chibby: Interplay between structural elements in target recognition. *Protein Sci.* **2016**, *25*, 1420–1429.

(55) Karunatilleke, N. C.; Fast, C. S.; Ngo, V.; Brickenden, A.; Duennwald, M. L.; Konermann, L.; Choy, W.-Y. Nrf2, the Major Regulator of the Cellular Oxidative Stress Response, is Partially Disordered. *Int. J. Mol. Sci.* **2021**, *22*, 7434.

(56) Gall, C.; Xu, H.; Brickenden, A.; Ai, X.; Choy, W. Y. The intrinsically disordered TC-1 interacts with Chibby via regions with high helical propensity. *Protein Sci.* **2007**, *16*, 2510–2518.

(57) Mokhtarzada, S.; Yu, C.; Brickenden, A.; Choy, W.-Y. Structural Characterization of Partially Disordered Human Chibby: Insights into Its Function in the Wnt-Signaling Pathway. *Biochemistry* **2011**, *50*, 715–726.

(58) Zahn, R.; Liu, A.; Luhrs, T.; Riek, R.; von Schroetter, C.; Garcia, F. L.; Billeter, M.; Calzolai, L.; Wider, G.; Wuthrich, K. NMR solution structure of the human prion protein. *PNAS* **2000**, *97*, 145–150.

(59) Wang, Y.; Fisher, J. C.; Mathew, R.; Ou, L.; Otieno, S.; Sublet, J.; Xiao, L.; Chen, J.; Roussel, M. F.; Kriwacki, R. W. Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21. *Nat. Chem. Biol.* **2011**, *7*, 214–221.

(60) Wong, L. E.; Kim, T. H.; Muhandiram, D. R.; Forman-Kay, J. D.; Kay, L. E. NMR Experiments for Studies of Dilute and Condensed Protein Phases: Application to the Phase-Separating Protein CAPRIN1. *J. Am. Chem. Soc.* **2020**, *142*, 2471–2489.

(61) Kim, D.-H.; Lee, J.; Mok, K.; Lee, J.; Han, K.-H. Salient Features of Monomeric Alpha-Synuclein Revealed by NMR Spectroscopy. *Biomolecules* **2020**, *10*, 428.

(62) Kosol, S.; Contreras-Martos, S.; Cedeño, C.; Tompa, P. Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy. *Molecules* **2013**, *18*, 10802–10828.

(63) Dyson, H. J.; Wright, P. E. NMR illuminates intrinsic disorder. *Curr. Opin. Struct. Biol.* **2021**, *70*, 44–52.

(64) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

(65) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **2012**, *134*, 3787–3791.

(66) Ahmed, M. C.; Skaanning, L. K.; Jussupow, A.; Newcombe, E. A.; Kragelund, B. B.; Camilloni, C.; Langkilde, A. E.; Lindorff-Larsen, K. Refine-

ment of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Front. Mol. Biosci.* **2021**, *8*, 216.

(67) Chang, M.; Wilson, C. J.; Karunatilleke, N. C.; Moselhy, M. H.; Karttunen, M.; Choy, W.-Y. Exploring the Conformational Landscape of the Neh4 and Neh5 Domains of Nrf2 Using Two Different Force Fields and Circular Dichroism. *J. Chem. Theory Comput.* **2021**, *17*, 3145–3156.

(68) Wilson, C. J.; Chang, M.; Karttunen, M.; Choy, W.-Y. KEAP1 Cancer Mutants: A Large-Scale Molecular Dynamics Study of Protein Stability. *Int. J. Mol. Sci.* **2021**, *22*, 5408.

(69) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513–5524.

(70) Cino, E. A.; Choy, W.-Y.; Karttunen, M. Characterization of the Free State Ensemble of the CoRNR Box Motif by Molecular Dynamics Simulations. *J. Phys. Chem. B* **2016**, *120*, 1060–1068.

(71) Samantray, S.; Yin, F.; Kav, B.; Strodel, B. Different Force Fields Give Rise to Different Amyloid Aggregation Pathways in Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2020**, *60*, 6462–6475.

(72) Nasica-Labouze, J.; Nguyen, P. H.; Sterpone, F.; Berthoumieu, O.; Buchete, N.-V.; Coté, S.; Simone, A. D.; Doig, A. J.; Faller, P.; Garcia, A.; Laio, A.; Li, M. S.; Melchionna, S.; Mousseau, N.; Mu, Y.; Paravastu, A.; Pasquali, S.; Rosenman, D. J.; Strodel, B.; Tarus, B.; Viles, J. H.; Zhang, T.; Wang, C.; Derreumaux, P. Amyloid β Protein and Alzheimer's Disease: When Computer Simulations Complement Experimental Studies. *Chem. Rev.* **2015**, *115*, 3518–3563.

(73) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *PNAS* **2013**, *110*, 5915–5920.

(74) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

(75) Best, R. B.; Hummer, G.; Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *PNAS* **2013**, *110*, 17874–17879.