HHS Public Access

Author manuscript

Curr Opin Struct Biol. Author manuscript; available in PMC 2023 November 24.

Published in final edited form as:

Curr Opin Struct Biol. 2022 June ; 74: 102372. doi:10.1016/j.sbi.2022.102372.

AlphaFold Illuminates Half of the Dark Human Proteins

Jessica L. Binder[1], **Joel Berendzen**[1,2], **Amy O. Stevens**[3], **Yi He**[1,3], **Jian Wang**[4], **Nikolay V. Dokholyan**[4,5], **Tudor I. Oprea**[1,6,7,8,*]

[1]Translational Informatics Division, Department of Internal Medicine, University of New Mexico, Albuquerque, NM 87131, USA.

[2]Current address: GenerisBio LLC, Santa Fe, NM 87507, USA.

[3]Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, NM 87131, USA.

[4]Department of Pharmacology, Department of Biochemistry and Molecular Biology, Penn State University College of Medicine, Hershey, PA 17033, USA.

[5]Department of Chemistry and Department of Biomedical Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, United States.

[6]UNM Comprehensive Cancer Center, Albuquerque, NM, USA.

[7]Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden.

[8]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

Abstract

We investigate the use of confidence scores to evaluate the accuracy of a given AlphaFold (AF2) protein model for drug discovery. Prediction of accuracy is improved by not considering confidence scores below 80 due to the effects of disorder. On a set of recent crystal structures, 95% are likely to have accurate folds. Conformational discordance in the training set has a much more significant effect on accuracy than sequence divergence. We propose criteria for models and residues that are possibly useful for virtual screening. Based on these criteria, AF2 provides models for half of understudied (dark) human proteins and two-thirds of residues in those models.

*Corresponding author: Oprea, T.I. | toprea@salud.unm.edu | Tel +15059254756.

Conflict of Interest

T.I.O. has received honoraria from or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth, and is on the Scientific Advisory Board of ChemDiv and InSilico Medicine.

Resources

We have implemented a $pLDDT_{80}$ classifier along with other useful features as a python package called *rafm*, which is installable via the usual mechanism of "*pip install rafm*" at the command line on systems with python 3.8 or greater installed. https://pypi.org/project/rafm

Supplementary Information contains pLDDT and derivative score information for the PDB subset, and for the Tdark subset of the human proteome, as well as AF2 model evaluation criteria.

INTRODUCTION

About half of Americans answering a 2020 survey would not get in an AI-driven taxi, and about three-quarters of them believed AI (artificial intelligence) cars were “not ready for primetime” [1]. Whether driving a vehicle or discovering new medicines, trust in AI depends on accumulated community experience and the consequences of errors in specific cases. There are over 20,000 protein-coding genes in the human genome [2-4]. Of these, 7074 have experimentally determined structures deposited in the Protein Data Bank (PDB) as of July 2021 [5]. Only 670 human proteins are therapeutically targeted by medicines, comprising the "drugged genome" [6]. Significant areas of biology remain potentially amenable to drug discovery [7]. Initiatives like “Illuminating the Druggable Genome” [8], OpenTargets [9], and Target 2035 [10] are exploring novel therapeutic opportunities in the “druggable” genome.

DeepMind described [11,12] AF2 (AlphaFold version 2.0), an AI method that predicts overall 3D structures of proteins. More than 350,000 AF2 structural models (including models of nearly every human protein) are now publicly accessible [12]. DeepMind garnered worldwide attention with their decisive win of the Critical Assessment of Techniques for Protein Structure Prediction, CASP14 [13]. Currently, scientists are assessing the impact of AF2 on research, including how much AF2 models expand the druggable genome.

Winning CASP14 presents a set of challenges specific to protein folding. However, protein 3D models do not often play a crucial role in drug discovery. The notion of *trust* in AF2 models is illustrated with a histogram of atomic Root-Mean-Square Deviations (*aRMSD*) on C_α atoms for crystal structures deposited in the PDB since AF2 was trained (*Fig. 2a* in [11]). It shows that AF2 produces high-quality folds in two-thirds of cases. However, the overall accuracy of a given AF2 model was not discussed. Local confidence scores (predicted Local Distance Difference Test, *pLDDT*) show a 95% per-residue correlation with experimentally-derived *LDDT* values [14] over the same proteins. AF2 model confidence evaluation is needed in the drug discovery context, given the non-local nature of *aRMSD*, the inherent selection bias of recent PDB structures, and the lack of any overall confidence-in-accuracy measure that can be calculated for individual models. Here, we discuss the issue of trust in AF2 models by addressing disorder, divergence, discordance, and druggability.

Disorder dominates confidence scores below 80

More than 30% of eukaryotic proteins contain one or more intrinsically disordered regions, IDRs [15-21]. Disorder is reflected in confidence scores as regions with low *pLDDT* [12]. Figure 1 displays the distributions of the *pLDDT* scores reported by AF2 for resolved/ordered and unresolved/disordered regions of crystal structures deposited in the PDB since AF2 was trained (the “post-AF2 test set”, see Supplemental Information). On this set of structures, ordered regions most frequently show *pLDDT* scores greater than 80, while IDRs have a broad distribution of *pLDDT* scores, with about 40% of unresolved regions falling below a *pLDDT* score of 50. From this analysis, we conclude that confidence scores below 80 are more indicative of disorder than of confidence in the accuracy of ordered structures, therefore in calculations on ordered-model accuracy we employ a cutoff of *pLDDT>80*.

Divergence has a minor effect on model accuracy

A problem with the 6-bin histogram used to estimate the distribution of model accuracies (*Fig. 2a* in [11]) is that *aRMSD* is a non-local measure. If a model is incorrect at the fold level, the expectation value of *aRMSD* scales with the radius of gyration. Thus, a model with 30 Å *aRMSD* against the experimental structure could be consistent with an entirely misfolded domain of around 1000 residues in length [22] or simply with rotation of a smaller domain about a single residue. To characterize different effects on model accuracy, we down selected the post-AF2 test set to 1,779 models that can be aligned with a corresponding experimental structure (see Supplemental Information) and used them to evaluate all-atom and backbone measures. *pLDDT* correlates poorly with log(*aRMSD*) on this set: Spearman rank correlation coefficient is 0.43 on the median (Supplemental Figure S1A). Truncating the range of *pLDDT* over which the median is calculated with a floor of 80 slightly improves the coefficient to −0.48 (Supplemental Figure S1B). We refer to the per-model median value of *pLDDT* scores greater than 80 as $pLDDT_{80}$.

Next, we split this down-selected test set into two pairs of subsets. The first pair explored high ($pLDDT_{80} > 90$) or low ($pLDDT_{80} < 88$) confidence scores. The second pair explored high (in clusters at 100% identity for over 80% of the length) or low (out of clusters at 5% identity for over 80% of the length) sequence identity to structures previously in the PDB. Cutoff values in these pairings were chosen to give maximal differences while maintaining roughly comparable fractions of the test set in the two arms of each pairing. We calculated distributions on log(*aRMSD*) and on the all-atom *LDDT* [14] for each of the subsets (Figure 2).

The *aRMSD* metric is not well-suited for characterizing structural models because its non-local nature tends to exaggerate the effects in small number backbone angle changes [23]. The lack of a high-difference tail in the low-identity *LDDT* distribution, together with the suppression of the high-difference tail in the low-confidence distribution, suggest that most differences between model and structure are in a few local coordinates, rather than many. Using *LDDT* as the accuracy measure improved Spearman's correlation on $pLDDT_{80}$ to 0.60 (Supplemental Figure S2).

Less than 1% of the high-confidence distribution appears in the range consistent with fold-level inaccuracies at $LDDT < 50$. Distributions for the low-similarity, high-confidence, and high-identity subsets are approximately the same, with the caveat afforded by the paucity of low-similarity models. But the 4% of models in the low-confidence distribution are distinctly worse than the other subsets. These observations suggest that AF2 produces models that are correct at the fold level more than 95% of the time, better than the previous two-thirds estimate [11]. The high-confidence and high-identity subsets similarity suggest that sequence divergence with PDB entries is not the primary driver of AF2 model inaccuracy for this set of structures.

Discordance limits model accuracy

Another driver of model inaccuracy is how AF2 handles differences among structures in the PDB with similar or identical sequences, a phenomenon we call *conformational discordance*.

AF2's training algorithm propagates conformational discordance through down-selecting among multiple PDB structures in a way that preserves maximum differences [11]. A key question is how well AF2 preserves correlations among conformational degrees of freedom–not just mean values and uncertainties–because those correlations are not needed to address the problem for which its algorithms were designed.

An illustrative case of conformational discordance is calmodulin (shown in Figure 3). Calmodulin is a kinase that, upon binding Ca^{++} ions, changes from a globular to a dumb-bell shape primarily through differences in two adjacent hinge residues [24]. In the calmodulin AF2 model, the effects of both conformations present in the training set are reflected in low confidence scores at the two hinge residues. While different AF2 runs yield slightly different results, none of the resulting models that we have sampled accurately reflect either the ion-free or ion-bound structures, but rather seem to be variations around the average of the two states and correlated changes in the two hinge residues have been lost. This example suggests that conformational discordance in the PDB results in *composite* AF2 models that rarely sample underlying conformations in the PDB.

Druggability: Are AF2 models ready for virtual screening?

Important structural elements relevant for drug discovery, such as prosthetic groups, ion binding, and protons are not included in AF2 models. Known protein conformational changes (as shown in Figure 3) can help us assess the effects of model accuracy on AF2 model readiness for target-based virtual screening (TBVS). If the model needs to be as close to the crystal structure as deoxy-myoglobin is to carboxy-myoglobin [25], only a tiny fraction of the AF2 models would be suitable for TBVS. If the model needs to be as close as R-state is to T-state hemoglobin, AF2 models may be suitable for characterizing allosteric sites [26]. A more typical TBVS example, where accuracy needs to be similar in capturing conformational changes, is when ERK2 is doubly phosphorylated. Given this example, a practical lower bound of global *pLDDT* of 80 could serve as basis for a model to likely be TBVS-ready. A value of *pLDDT* of 80 indicates a 68% likelihood of sidechain rotamers falling into the correct hemisphere (Fig 2B in [11]). Surfaces formed by two adjacent residues with $pLDDT \geq 80$ are very close to the 50% accuracy limit if rotamer errors are independent. Having previously introduced $pLDDT_{80}$, we set $pLDDT_{80} \geq 91.2$ as criterion for assessing AF2 model quality, combined with the fraction of protein length for which this holds true ($pLDDT_{80}$_frac) to evaluate TBVS potential; see Supplemental Information. These criteria allow us to calculate a confusion matrix (see Supplemental Figure S2) that gives the sensitivity (true positive rate of classification) of 90.1% and a precision (positive predictive value of classification) of 86.3%.

Given these criteria, we evaluated which AF2 models of the human understudied proteins, Tdark [7], which currently lack an experimental PDB structure might be TBVS-ready (Figure 4A). Of the set of 5592 “dark” proteins with AF2 models, 3051 (54.6%) meet our criteria for possibly being accurate enough for TBVS studies (Figure 4B). Taking into account the estimated false-positive rate (~6% of total), this implies that AF2 provides TBVS-ready models for about half of the understudied human proteins. Additionally, among the Tdark proteins associated with very high or confident AF2 models, 664 match druggable

protein classes as follows: 235 enzymes; 23 GPCRs; 220 immune response proteins; 5 ion channels; 16 kinases; 45 receptors; 25 signaling proteins; 97 transporters. More strictly, by matching Tclin protein motif domains (according to PFAM, InterPro, and Prosite) with Tdark proteins, 32 of the above 664 may be more likely to have “druggable” binding domains[27]. In total, 50 Tclin associated motifs are present in “very high” or “confident” Tdark AF2 models; see Supplemental Information.

CONCLUSIONS

In our opinion, future work would do well to move away from the familiar aRMSD metric of overall model-structure agreement in favor of LDDT or other local measures. The aRMSD metric suggests that AF2 models are worse compared to pLDDT. Structural bioinformatics would also benefit from developing measures that disambiguate the effects of disorder, discordance, and divergence.

Proteins can take on different conformations, and which protein conformation is more druggable depends on the clinical need associated with a particular disease state. Screening the right target in the wrong conformation reduces the likelihood of finding valuable leads. A few proteins are represented in the PDB by structures determined in multiple conformations, while most have only one. Many of the biophysical drivers that determine protein conformation, such as the hydrophobic effect, are poorly understood at present. More work will be needed on the taxonomy of possible protein conformations before AI approaches can be expected to tackle the problem of conformation robustly. Until then, it might be helpful to eliminate discordant structures from the training set to predict a single conformation with higher accuracy than achieved with composite conformations at present.

AF2 forces us to reconsider the implications of disorder on druggability because it performs well at predicting IDRs [28]. Having trust that a region is disordered versus trusting the ordered region's accuracy leads to different conclusions. It is worth noting that the dataset used in this work primarily includes proteins enriched with relatively short IDRs, as 95% of proteins in the dataset are at most 29% unresolved/disordered. In a scenario where the dataset included proteins with significantly longer stretches of IDRs, the results from current work may not apply. Proteins containing IDRs play critical roles in many biological functions [29-36] and are associated with various diseases [37-40]. Thus, IDRs are potential targets in drug discovery [41-43]. “Disordered” does not mean “undruggable” because unique strategies for drug discovery in targets containing disordered regions are available [44]. Regions of pLDDT < 50 in an AF2 model indicate those strategies could be employed. Moreover, the existence of a structural model is neither necessary nor a sufficient condition for drug discovery. Even the use of high-quality experimental structures of the correct conformational state does not guarantee successful TBVS hits.

About 5% of the human "dark" proteome has structures in the PDB (Supplemental Information). Cost-to-benefit analyses of whether to deploy TBVS on AF2 models remain project-dependent. However, AF2 model quality may be "good enough" for rapid deployment for over 3000 understudied human proteins. AF2 models may help de-risk protein targets through protein expression and solubility and may provide protein

engineering suggestions. By identifying likely boundaries of compact domains, disordered regions, or linkers, AF2 and other methods can enable synthesis-by-domain strategies that can break large proteins into more tractable modules to be expressed or synthesized then reconstituted in-vitro. Regardless of its impact on *in silico* technologies, AF2 does not preclude structural biology and structure-based drug design. However, AF2 is poised to become a powerful tool in the evolving drug discovery arsenal.

Computational models are very different from experimental structures in that they can be updated on-demand with the latest improvements. Public notebooks such as ColabFold [45] facilitate the removal of disordered termini, improving sequence alignment, adding a binding partner, and calculating new models within minutes. Although not designed with protein oligomers or assemblies in mind, multiple groups are working on use of AF2 to illuminate protein-protein interactions. In 2014, it was estimated that 40% of protein structures were experimentally determined [46]. With AF2 and future improvements, structural biology, and drug discovery are about to exponentially increase with new computational tools that combine sequence evolution, structures, and ligand binding knowledge [47].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge support by the National Institutes for Health IDG KMC application from the University of New Mexico (Grant No. U24CA224370), the Passan Foundation (Grant No. R35 GM134864 to N.V.D.), the National Science Foundation (Grant No. 2137558 to Y.H.), and the National Science Foundation Graduate Research Fellowship (Grant No. DGE-1939267).

REFERENCES

1. PAVE Poll: Americans wary of AVs but say education and experience with technology can build trust. 18 May 2020 [cited 23 Sep 2021]. Available: https://pavecampaign.org/pave-poll-americans-wary-of-avs-but-say-education-and-experience-with-technology-can-build-trust/
2. National Center for Biotechnology Information. [cited 16 Sep 2021]. Available: https://www.ncbi.nlm.nih.gov/
3. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2007;35: D26–31. [PubMed: 17148475]
4. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019;12: 315. [PubMed: 31164174]
5. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Research. 2021. pp. D437–D451. doi:10.1093/nar/gkaa1038 [PubMed: 33211854]
6. Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen D-T, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res. 2021;49: D1160–D1169. [PubMed: 33151287] DrugCentral is a public portal that provides up-to-date drug information. The current release includes newly approved active pharmaceutical ingredients (current through June 2021), pharmacokinetic properties for ~1000 drugs, sex-based separation of side effects processed from FAERS (FDA Adverse Event Reporting System), and a machine learning platform that estimates anti-SARS-CoV-2 activities, REDIAL-2020.

7. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nature Reviews Drug Discovery. 2018;17: 317–332. [PubMed: 29472638]
8. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen D-T, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res. 2021;49: D1334–D1346. [PubMed: 33156327] The Target Central Resource Database (TCRD) is an open-access resource developed as a part of the IDG program and jointly serves as the knowledge hub for over 20 000 human protein targets (collating information from several gene/protein data sources). And Pharos is the web interface to browse the TCRD. Pharos recently added AF2 structures to each protein target and continues to add useful information to empower users to find new areas of study in the druggable genome.
9. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: supporting systematic drug-target identification and prioritisation. Nucleic Acids Res. 2021;49: D1302–D1310. [PubMed: 33196847]
10. Carter AJ, Kraemer O, Zwick M, Mueller-Fahrnow A, Arrowsmith CH, Edwards AM. Target 2035: probing the human proteome. Drug Discov Today. 2019;24: 2111–2115. [PubMed: 31278990]
11. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. [PubMed: 34265844] The authors describe the successful development/methods of AF2, including the incorporation of evolutionary information through multiple-sequence alignments and the use of transformers to handle rotational and translational symmetries in an equivariant-attention fashion. Based on early citations, it appears this paper is likely to become one of the most cited papers in contemporary science. Of particular interest is Fig. 2A of this paper which shows the distribution of mean atomic root-mean-square deviations between models and a subset (those that did not cluster with PDBclust) of crystal structures deposited since AF2 was trained.
12. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596: 590–596. [PubMed: 34293799] This paper documents the deposition of 350,000 structural models produced by AlphaFold2, including models for 99% of the human genome. Of particular interest is the discussion and how disordered regions in proteins correlate with confidence scores.
13. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins. 2021. doi:10.1002/prot.26257
14. 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. [PubMed: 23986568] This paper introduces the local distance difference test (LDDT), a superposition-free score that combines an agreement-based model quality measure with stereochemical plausibility checks. LDDT serves as tool for evaluating protein structure predictions.
15. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci. 2015;72: 137–151. [PubMed: 24939692]
16. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11: 161–171.
17. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn. 2012;30: 137–149. [PubMed: 22702725]
18. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004;337: 635–645. [PubMed: 15019783]
19. Dunker AK, Keith Dunker A, David Lawson J, Brown CJ, Williams RM, Romero P, et al. Intrinsically disordered protein. Journal of Molecular Graphics and Modelling. 2001. pp. 26–59. doi:10.1016/s1093-3263(00)00138-8 [PubMed: 11381529]
20. Uversky VN, Gillespie JR, Fink AL. Why are ?natively unfolded? proteins unstructured under physiologic conditions? Proteins. 2000;41: 415–427. [PubMed: 11025552]

21. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293: 321–331. [PubMed: 10550212]
22. Tanner JJ. Empirical power laws for the radii of gyration of protein oligomers. Acta Crystallogr D Struct Biol. 2016;72: 1119–1129. [PubMed: 27710933]
23. Kufareva I, Abagyan R. Methods of protein structure comparison. Methods Mol Biol. 2012;857: 231–257. [PubMed: 22323224]
24. Zhang M, Tanaka T, Ikura M. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. Nat Struct Biol. 1995;2: 758–767. [PubMed: 7552747]
25. Vojtechovský J, Chu K, Berendzen J, Sweet RM, Schlichting I. Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophys J. 1999;77: 2153–2174. [PubMed: 10512835]
26. Safo MK, Ahmed MH, Ghatge MS, Boyiri T. Hemoglobin–ligand binding: Understanding Hb function and allostery on atomic level. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2011;1814: 797–809. [PubMed: 21396487]
27. Lin Y, Mehta S, Küçük-McGinty H, Turner JP, Vidovic D, Forlin M, et al. Drug target ontology to classify and integrate drug discovery data. J Biomed Semantics. 2017;8: 50. [PubMed: 29122012]
28. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci. 2017;74: 3069–3090. [PubMed: 28589442]
29. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit. 2005;18: 343–384. [PubMed: 16094605]
30. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008;47: 7598–7609. [PubMed: 18627125]
31. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Keith Dunker A, Obradovic Z, et al. Functional Anthology of Intrinsic Disorder. 2. Cellular Components, Domains, Technical Terms, Developmental Processes, and Coding Sequence Diversities Correlated with Long Disordered Regions. Journal of Proteome Research. 2007. pp. 1899–1916. doi:10.1021/pr060393m [PubMed: 17391015]
32. Fuxreiter M, Tompa P, Simon I, Uversky VN, Hansen JC, Asturias FJ. Malleable machines take shape in eukaryotic transcriptional regulation. Nat Chem Biol. 2008;4: 728–737. [PubMed: 19008886]
33. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev. 2014;114: 6561–6588. [PubMed: 24739139]
34. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. Biochemistry. 2006;45: 6873–6888. [PubMed: 16734424]
35. Deiana A, Forcelloni S, Porrello A, Giansanti A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. PLoS One. 2019;14: e0217889. [PubMed: 31425549]
36. Bondos SE, Dunker AK, Uversky VN. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. Cell Commun Signal. 2021;19: 88. [PubMed: 34461937]
37. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys. 2008;37: 215–246. [PubMed: 18573080]
38. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome. Protein Pept Lett. 2009;16: 1533–1547. [PubMed: 20001916]
39. Kulkarni P, Uversky VN. Intrinsically Disordered Proteins in Chronic Diseases. Biomolecules. 2019;9. doi:10.3390/biom9040147
40. Coskuner O, Uversky VN. Intrinsically disordered proteins in various hypotheses on the pathogenesis of Alzheimer's and Parkinson's diseases. Progress in Molecular Biology and Translational Science. 2019. pp. 145–223. doi:10.1016/bs.pmbts.2019.05.007 [PubMed: 31521231]
41. Ruan H, Sun Q, Zhang W, Liu Y, Lai L. Targeting intrinsically disordered proteins at the edge of chaos. Drug Discov Today. 2019;24: 217–227. [PubMed: 30278223]

42. Santofimia-Castaño P, Rizzuti B, Xia Y, Abian O, Peng L, Velázquez-Campoy A, et al. Targeting intrinsically disordered proteins involved in cancer. Cellular and Molecular Life Sciences. 2020. pp. 1695–1707. doi:10.1007/s00018-019-03347-3 [PubMed: 31667555]
43. Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol. 2021; 167208. The authors review the top predictors of intrinsically disordered regions (IDRs) while also highlighting the future expansions in protein structure prediction led by machine learning techniques, including AF2. While there exists a large overlap between low AF2 confidence scores and experimentally determined IDRs, they note that a cautionary approach must be taken before assuming a concrete understanding of protein structure.
44. Tsafou K, Tiwari PB, Forman-Kay JD, Metallo SJ, Toretsky JA. Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm. Journal of Molecular Biology. 2018. pp. 2321–2341. doi:10.1016/j.jmb.2018.04.008 [PubMed: 29655986]
45. Sergey O. ColabFold: Making Protein folding accessible to all via Google Colab! Github; Available: https://github.com/sokrypton/ColabFold ColabFold is a repository of notebooks that make AlphaFold2 and other AI-driven structural modeling methods available to users using the latest developments. Of particular interest are the development of methods for running models employing homooligomers, custom multiple-sequence alignments, and assemblies.
46. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A. 2014;111: 3733–3738. [PubMed: 24567391]
47. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med. 2021;27: 1666–1669. [PubMed: 34642488] The authors comment on the variable quality of AF2 models (“good, bad and ugly”) and distributions of per-model average pLDDT scores across four model organisms: human, Trypanosoma cruzi, Mycobacterium tuberculosis and Escherichia coli. They note that M. tuberculosis and E. coli have twice as many “very high” confidence scores compared to those in and T. cruzi, due to differences in average sequence lengths. Given our findings on the large effects that disorder has on untruncated confidence scores, these differences could also reflect higher amounts of disordered regions in eukaryotes compared with bacteria.
48. Database APS. AlphaFold Protein Structure Database. [cited 28 Sep 2021]. Available: https://alphafold.ebi.ac.uk
49. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. 2016;32: 1323–1330. [PubMed: 26743509]

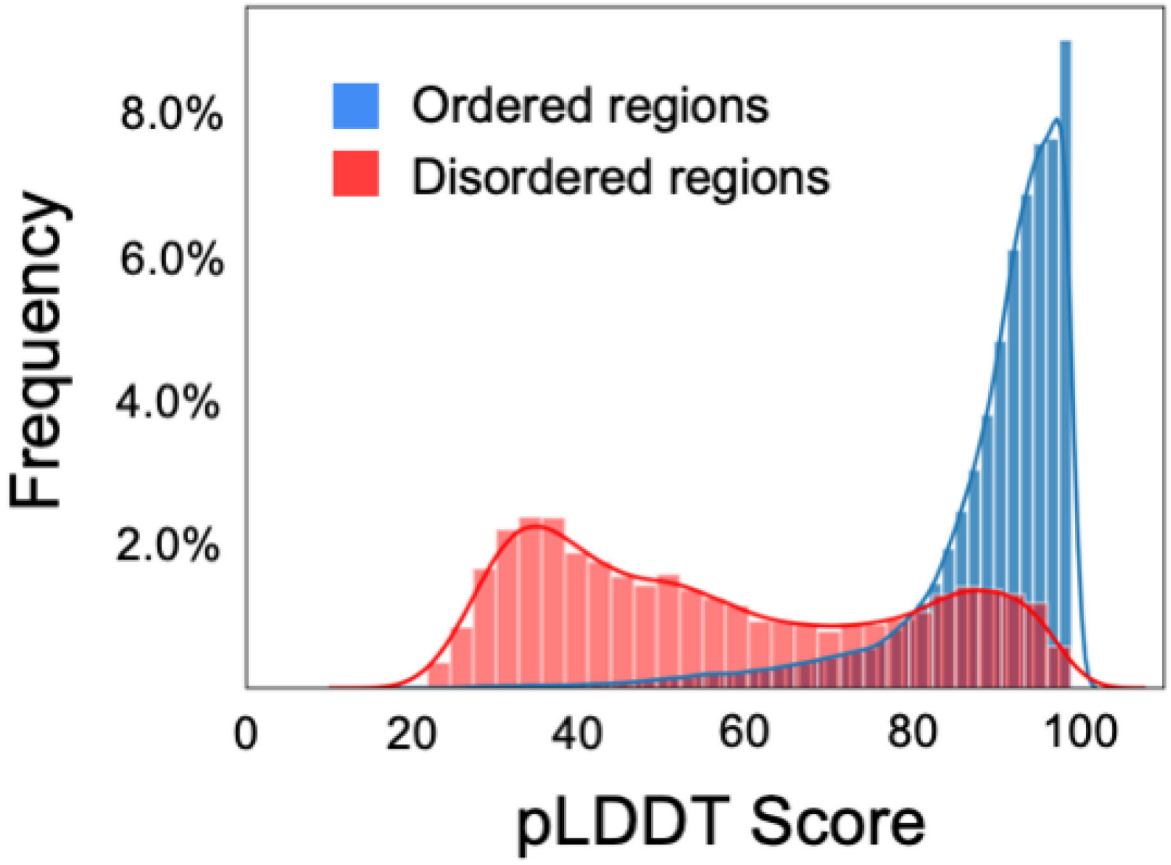

Figure 1. Distribution of AlphaFold confidence scores across ordered (blue) and disordered (red) regions.

Ordered and disordered regions correspond to resolved and unresolved parts, respectively, for the post-AF2 test set. Terminal regions were not included. Ordered regions most frequently show pLDDT *scores >80%. Disordered regions show a broad distribution of* pLDDT *scores with comparable frequencies from* pLDDT *scores between 20% and 90%.*

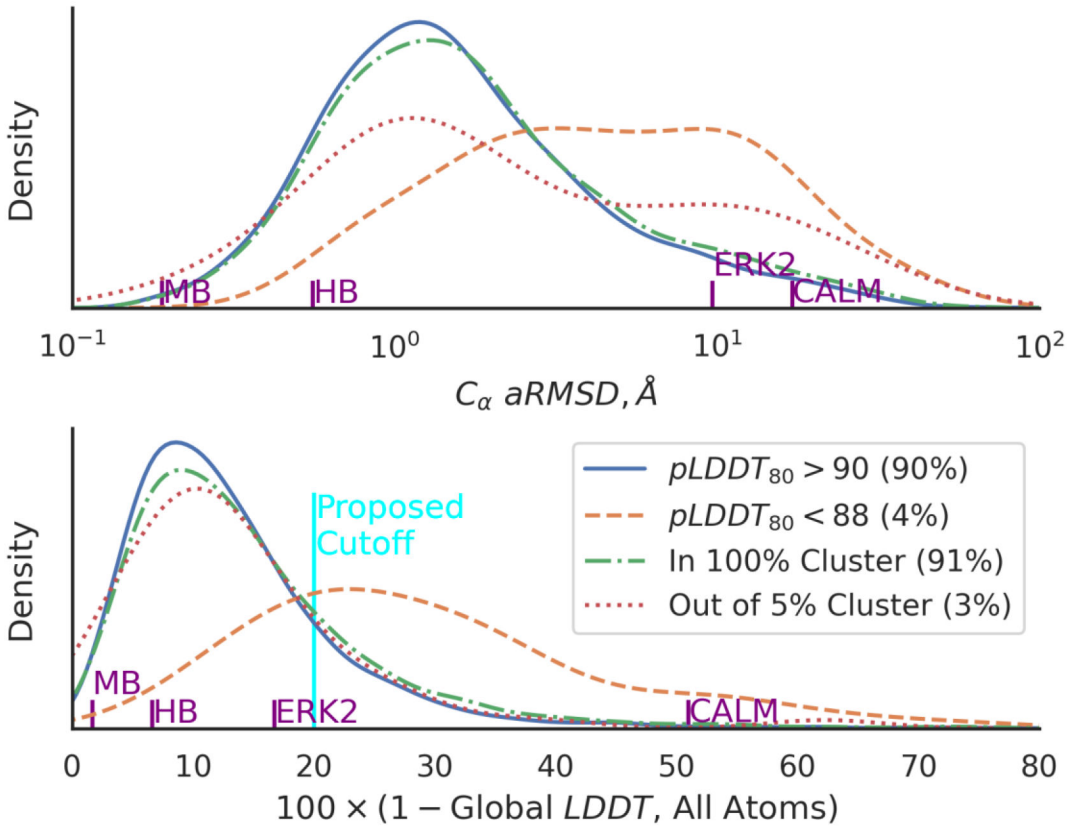

Figure 2. Kernel-density estimates of the distribution of differences between AF2 models and crystal structures using atomic Root-Mean-Square Displacements on C_α atoms (top) and Global Local Distance Difference Test metrics (bottom).

These distributions were calculated from crystallographic structures that were deposited in the Protein Data Bank after the AF2 training set cut-off date of April 30, 2018. Only residues that unambiguously intersect between AF2 models deposited in EMBL [48] and crystal structures were considered, with a minimum per-chain length cutoff of 20 residues, resulting in 1810 structural models to be compared. Distributions for mean confidence levels ($pLDDT_{80}$) over the raw models at or above 90 (solid blue line) and below 88 (dotted orange line) are plotted. We also clustered the PDB using mmseqs [49] to select for sequences nearly identical to an existing structure (in clusters with 100% minimum sequence identity over 80% of the longest sequence and cluster mode 2, dot-dash green line) or decisively non-matching regions (out of 5% minimum sequence identity, dotted red line). The high-confidence ($pLDDT_{80}$>90) distribution on log(aRMSD) peaks at 1.7 Å aRMSD, with a long tail extending beyond 10 Å at the 10% level. The low-confidence distribution on log(aRMSD) has a broad flat shape suggesting peaks at 3 and 20 Å. The high-identity distribution looks similar to the high-confidence distribution, while the low-identity distribution has peaks near 2 and 20 Å, respectively. Plotted against all-atom LDDT, the high-confidence, high-identity, and low-identity distributions look similar to each other; only the low-confidence distribution is distinct, with a single peak at LDDT ~75. Notations on the x-axis indicate differences between structures of ligand-free vs. ligand bound myoglobin (MB, PDB entries 1A6N and 1A6G); R- vs. T-state hemoglobin (HB, PDB entries 6BWP and 6BWU); unphosphorylated vs. doubly-phosphorylated conformations

of an extracellular signal-regulated kinase (ERK2, PDB entries 1ERK and 2ERK0); and calcium-free vs. calcium-bound calmodulin (CALM, PDB entries 1CLL and 1QX5). The cyan line shows the proposed LDDT cutoff for a structure that is likely to be useful for virtual screening.

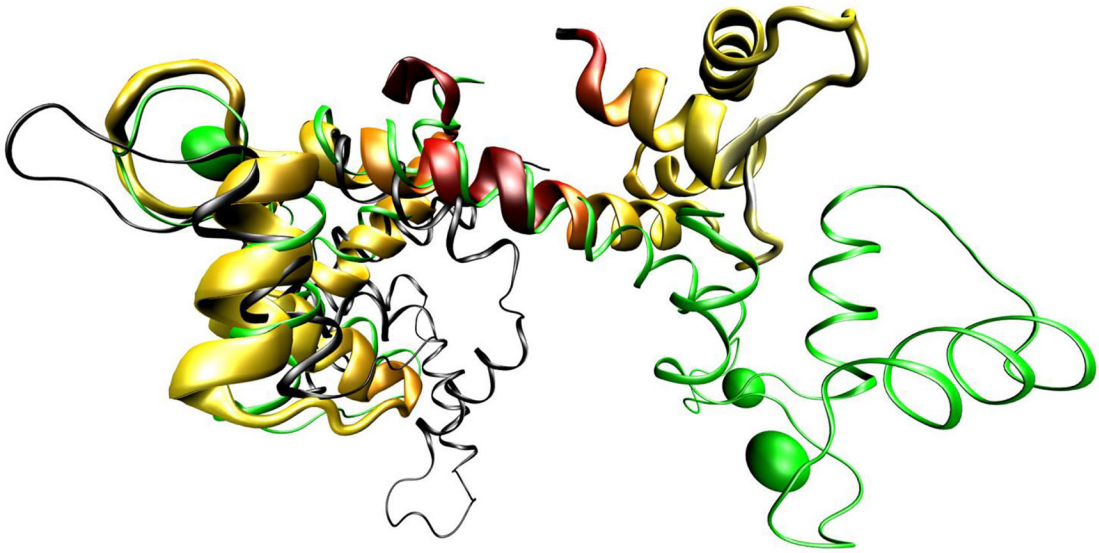

Figure 3. Comparison of discordant crystal structures of calmodulin with an AF model.

The calcium-bound crystal structure (PDB entry 1CLL, thin green cartoon with Ca^{++} *ions as spheres), with alignments against the first half of the calcium-free crystal structure (PDB entry 1QX5, thin black cartoon) and the AlphaFold2 model (P0DP23-F1-model_v1, thick yellow-red cartoon), aligned on their N-terminal halves. Yellow regions of the model represent very high confidence (*pLDDT *> 90) residues, while dark-red regions represent very low confidence (*pLDDT *<50) residues. The low confidence region at the center of the AF model corresponds to a hinge where the calcium-bound and calcium-free models diverge. When aligned in this manner, aRMSD values of 7.2 Å against the calcium-bound structure and 6.7 Å against the calcium-free structure were obtained. When aligned across all residues, the AF model yields aRMSDs of 10 Å against the calcium-bound structure and 17 Å against the calcium-free structure, respectively. Global* LDDT *scores for the experimental structures are 49% for all atoms and 56% for* C_α *only.*

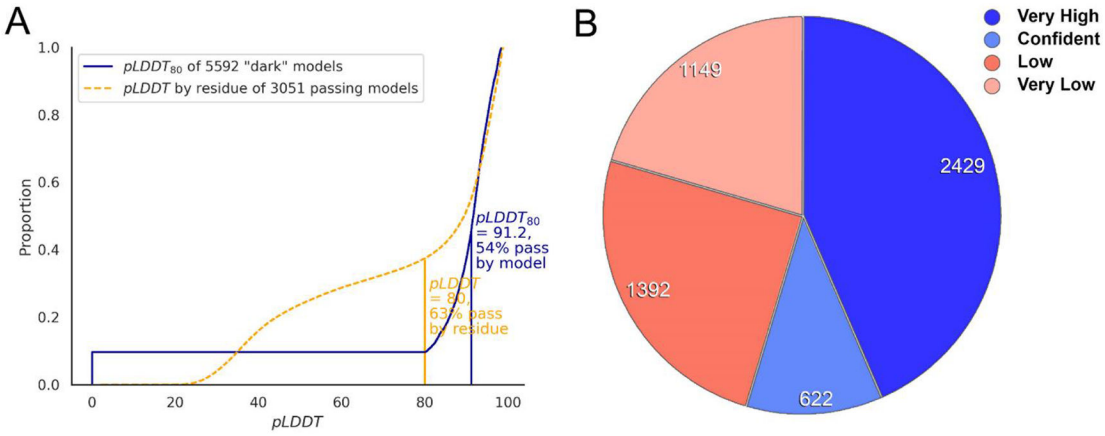

Figure 4. Fraction of the dark genome potentially illuminated by AF2 models.

A) *Of the set of 5592 unique "dark" proteins with AF2 models, 3051 (54%) pass the proposed selection criteria of pLDDT80 greater to or equal to 91.2 while having at least 20 residues with pLDDT ≥ 80.* **B)** *Pie chart illustrating AF2 model quality according to pLDDT80-derived criteria (see* Supplementary Information*): 3051 (54%) proteins associated with "very high" or "confident" AF2 models are likely to be TBVS-ready, whereas 2541 proteins are not.*