# SUPPLEMENTARY INFORMATION FOR:
Metapredict enables accurate disorder prediction across the Tree of Life

Jeffrey M. Lotthammer[1, 2], Jorge Hernández-García[3], Daniel Griffith[1, 2], Dolf Weijers[3], Alex S. Holehouse[1, 2, 4], Ryan J. Emenecker[1,2,*]

1. Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, 63100, USA
2. Center for Biomolecular Condensates, Washington University, St. Louis, MO 63130, USA
3. Laboratory of Biochemistry, Wageningen University, Stippeneng 4, 6708WE Wageningen, The Netherlands
4. Lead contact

*Correspondence: remenecker@wustl.edu, alex.holehouse@wustl.edu
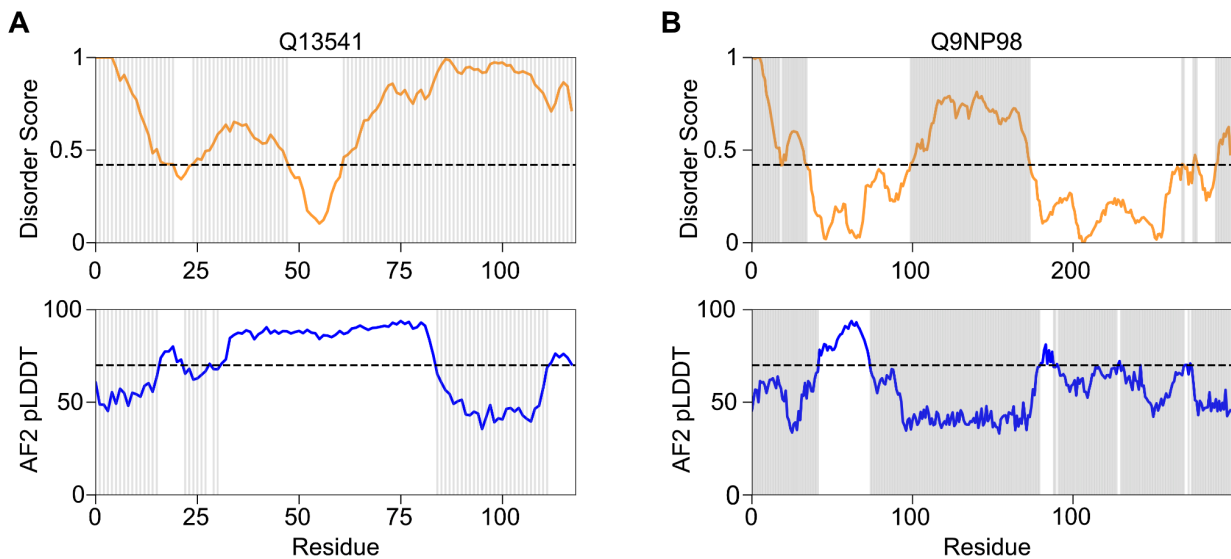
Supplemental Fig. 1. **AF2 pLDDT scores and metapredict V1 disorder scores identify different protein regions as disordered.**
(A) Disorder scores were generated using the metapredict V1 network (top) and AF2 pLDDT values (bottom) for *Homo sapiens* EIF4EBP1 (Uniprot Q13541). For disorder scores (top), regions highlighted are above the threshold value of 0.42 and thus are classified as disordered. For AF2 pLDDT values (bottom), scores below 70 (classified as low or very low confidence by AF2) are highlighted.
(B) Same as for (A) except for *Homo sapiens* Myozenin-1 (Uniprot Q9NP98).
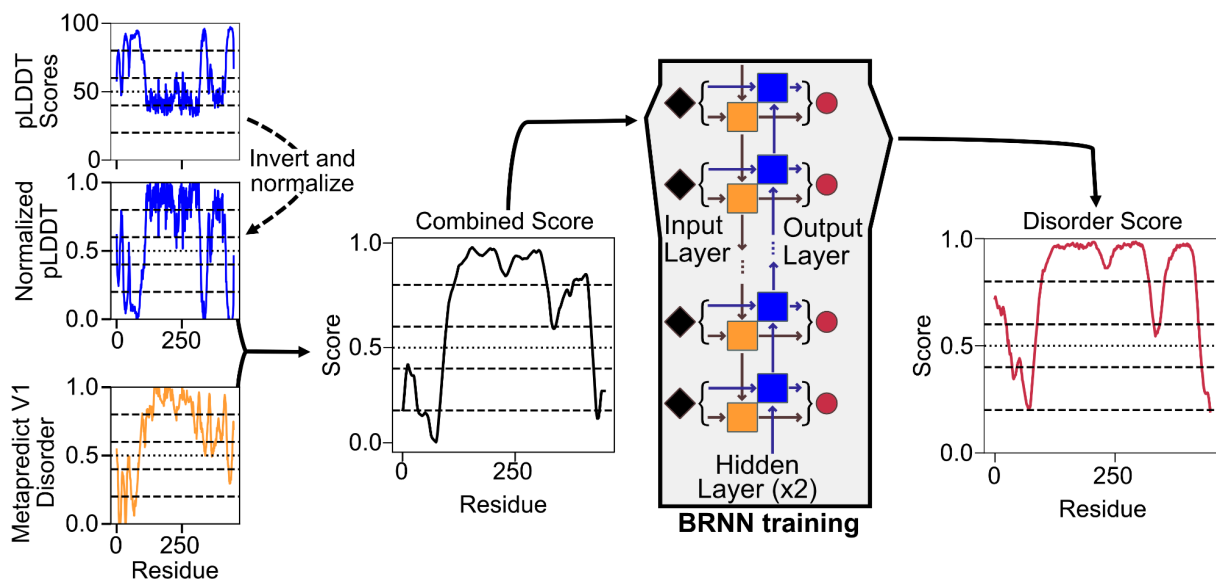
Supplemental Fig. 2. Generating combined disorder scores from metapredict V1 disorder scores and AF2 pLDDT.
The approach for generating the 'combined disorder scores' for metapredict V3 was as follows: (1) pLDDT scores (top left) were divided by 100 and then were normalized using a base value equal to 0.35 and a top value equal to 0.95 such that final values would be between 0 and 1 and then were inverted such that high scores correspond to disorder and low scores correspond to not disordered (left side, middle), (2) metapredict V1 disorder scores (left side, bottom) were combined with pLDDT scores such that if either value was greater than 0.5, that value was used, and if both values were below 0.5, the lower value was used. A Savitzky-Golay filter was then applied to these scores, and scores above 1 were set to 1 and scores below 0 were set to 0. Finally, a smoothing step using a moving average with a window size of 25 was applied. These combined scores (graph labeled 'combined') were then generated for 455,666 protein sequences, and the resulting scores were used to train a BRNN-LSTM. The scores generated by the resultant network (furthest right graph) accurately recapitulate the 'combined disorder scores.'
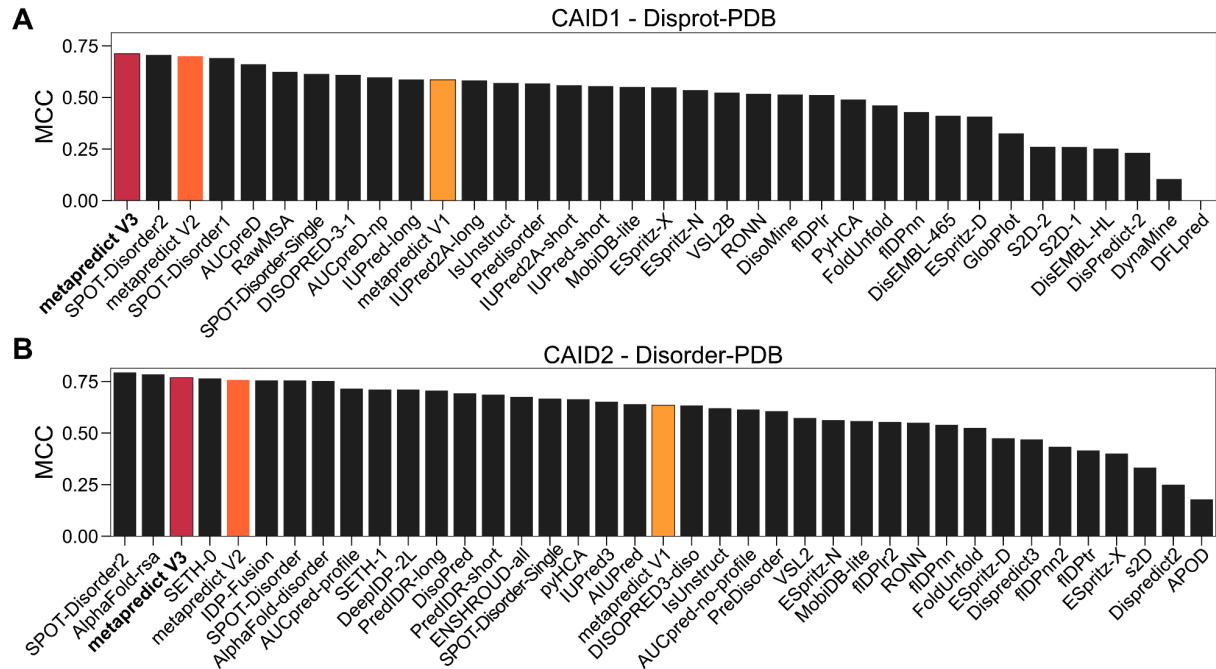
Supplemental Fig. 3. CAID1 and CAID2 disorder assessment using MCC to quantify accuracy for all three available metapredict networks.

(A) Matthew's correlation coefficient (MCC) of various disorder predictors tested for accuracy using the Disprot-PDB dataset from CAID1 [1]. MCC values shown for predictors other than metapredict V2 and V3 are directly from the CAID1 dataset [1].

(B) Same as (A) except using data from the CAID2 analysis, which utilized a new dataset and included additional disorder predictors [2]. For the CAID2 MCC values, binarized values of disorder prediction were required to calculate MCC, so only disorder predictors that provided a 'disorder cutoff value' to CAID or those with a publicly available disorder cutoff threshold are included in Fig. 1B.
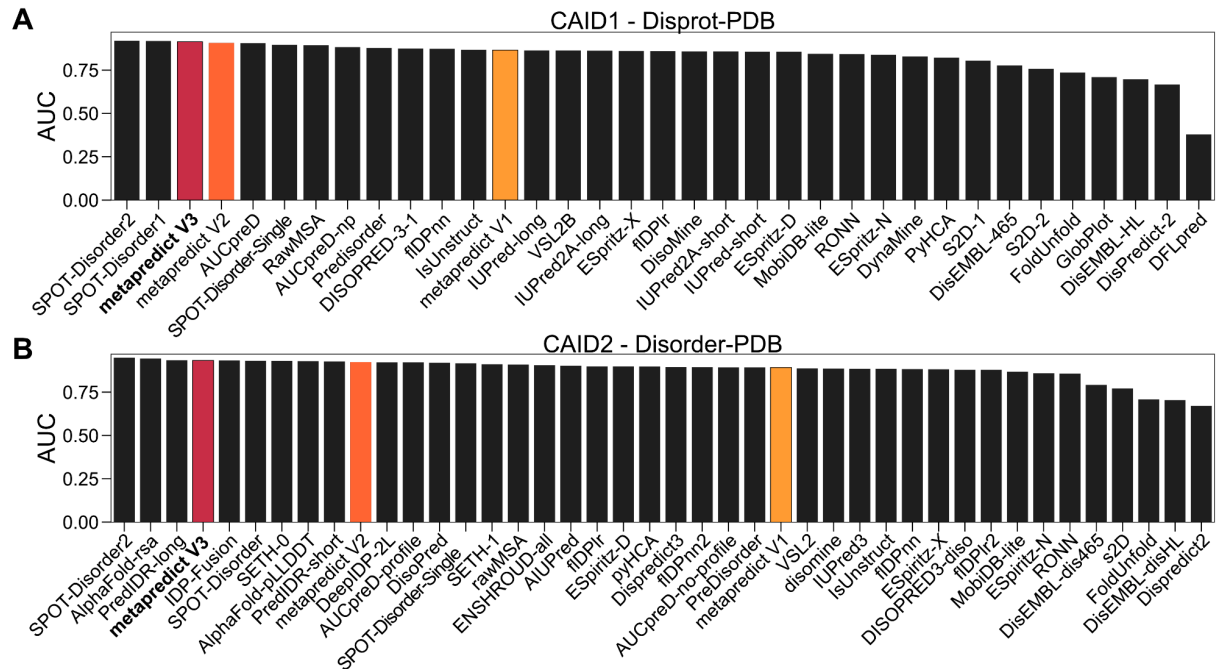
Supplemental Fig. 4. CAID1 and CAID2 disorder assessment using AUC to quantify accuracy.

(A) Area under the receiver-operating characteristic curve (AUC) of various disorder predictors tested for accuracy using the Disprot-PDB dataset from CAID1 [1]. AUC values shown for predictors other than metapredict V2 and V3 are directly from the CAID1 dataset [1].

(B) Same as (A) except using data from the CAID2 analysis, which utilized a new dataset and included additional disorder predictors [2].
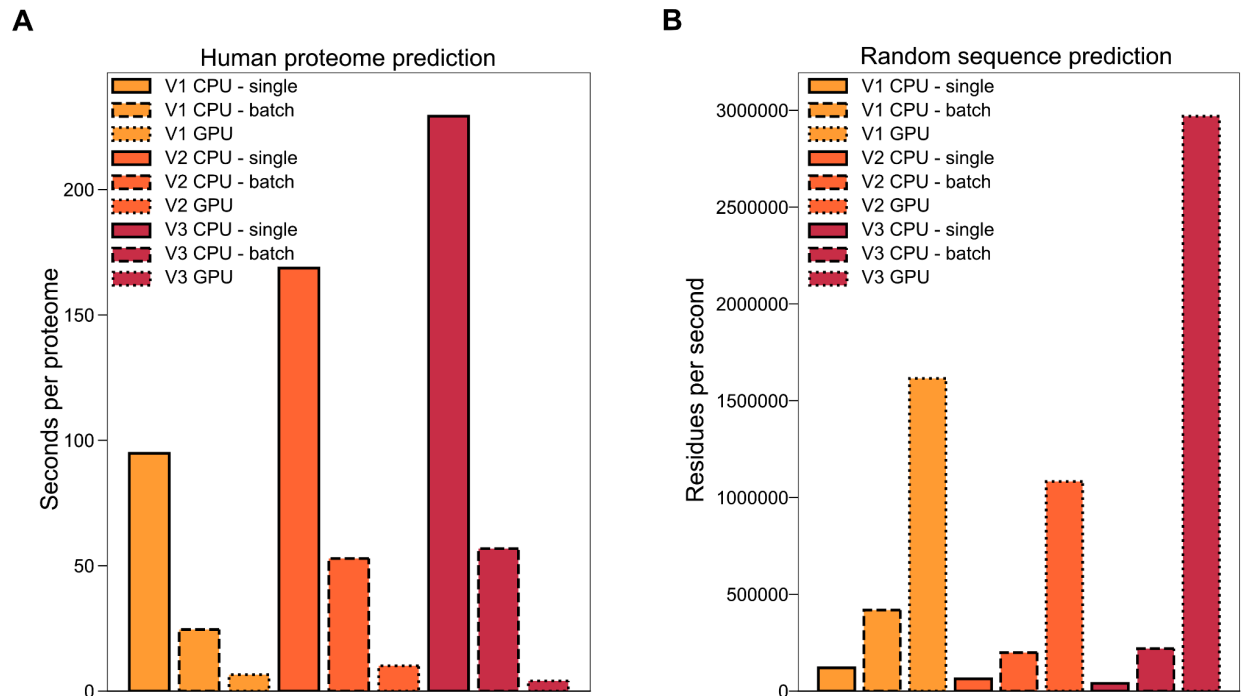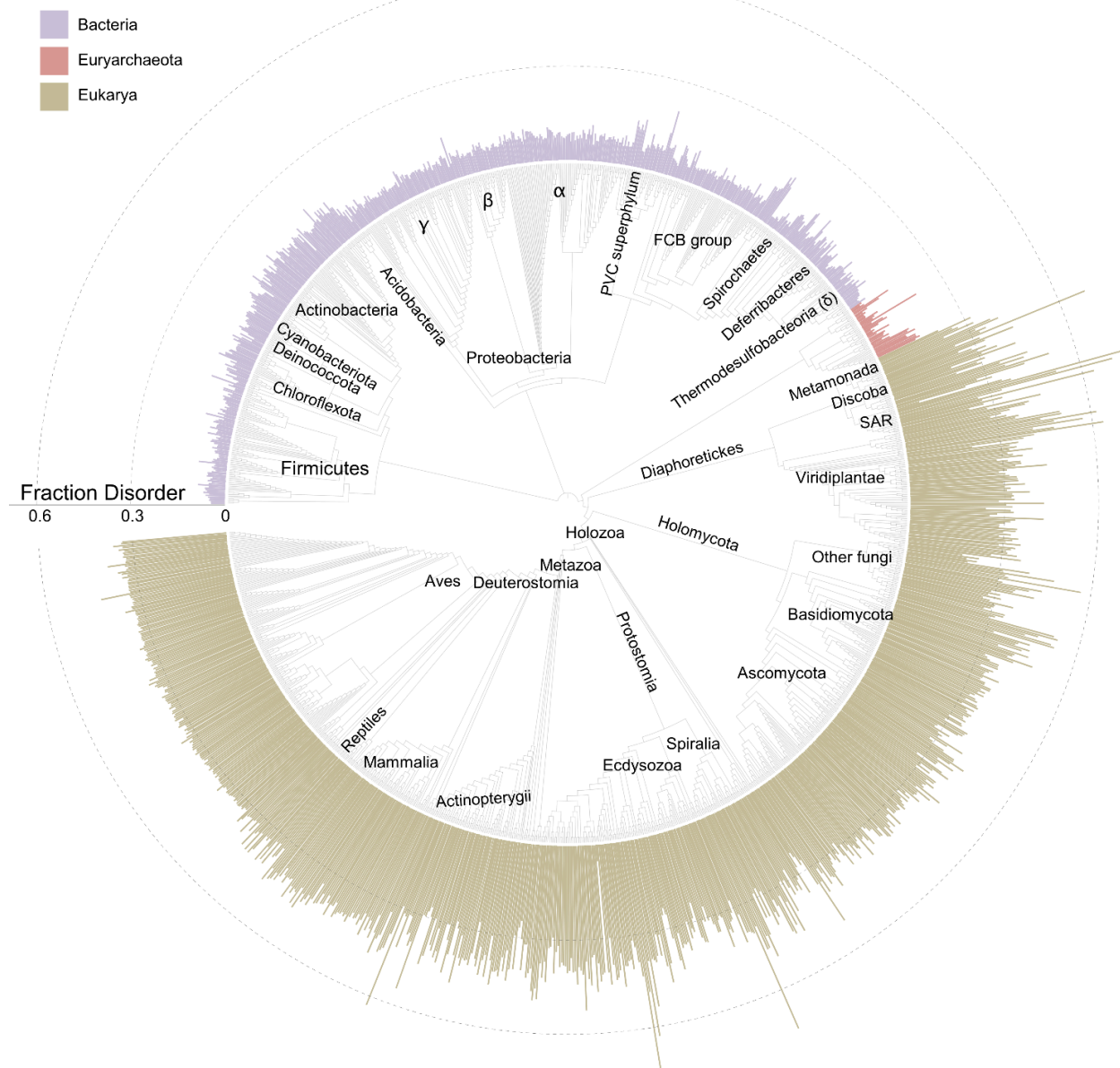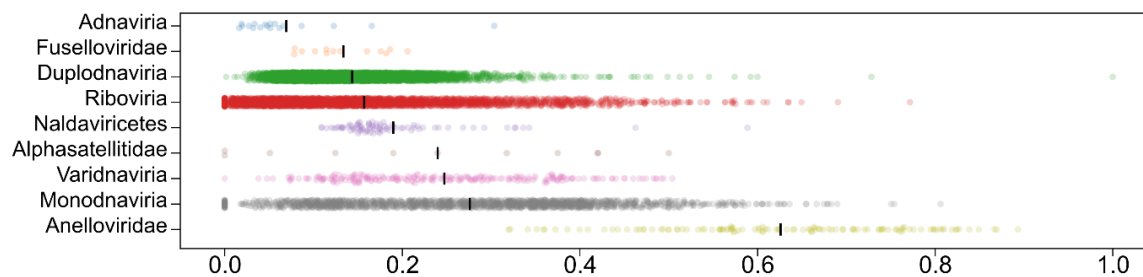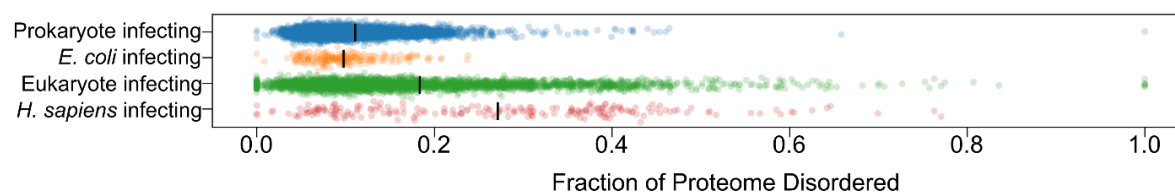
Supplemental Fig. 5. Comparing prediction speeds between all three metapredict networks.

The speed at which metapredict generates disorder predictions depends on the network used. All three networks are available for use in our implementation of metapredict V3. Using the architecture from metapredict V3, including improvements in the backend code, we tested the speed for disorder prediction using each of our networks. Notably, the V3 network carries out predictions using a larger batch size than V1 or V2 and thus shows a greater improvement in prediction speed when predictions are carried out in batch and when carried out using a GPU.

(A) Prediction speed in seconds for the different versions of metapredict using CPU without batch prediction enabled, CPU with batch prediction, and prediction using GPU. As with Figure 1D, the reference human proteome from UniProt (one protein sequence per gene, 20,594 proteins in total) was used.

(B) Prediction speed in residues per second for the different versions of metapredict using CPU without batch prediction enabled, CPU with batch prediction, and prediction using GPU. Prediction speed was calculated by dividing the number of seconds required to predict the disorder for 500,000 randomly generated sequences by the sum of the total number of residues in the 500,000 sequences.

A

Bacteria
Euryarchaeota
Eukarya

γ β α PVC superphylum FCB group
Acidobacteria Proteobacteria Spirochaetes
Actinobacteria Deferribacteres
Cyanobacteriota Thermodesulfobacteria (δ)
Deinococcota Metamonada
Chloroflexota Discoba
SAR
Firmicutes Diaphoretickes Viridiplantae

Fraction Disorder
0.6 0.3 0

Holozoa Holomycota Other fungi
Metazoa Basidiomycota
Aves Deuterostomia Ascomycota
Protostomia
Reptiles
Mammalia Spiralia
Ecdysozoa
Actinopterygii

B

Adnaviria
Fuselloviridae
Duplodnaviria
Riboviria
Naldaviricetes
Alphasatellitidae
Varidnaviria
Monodnaviria
Anelloviridae

0.0   0.2   0.4   0.6   0.8   1.0

C

Prokaryote infecting
E. coli infecting
Eukaryote infecting
H. sapiens infecting

0.0   0.2   0.4   0.6   0.8   1.0

Fraction of Proteome Disordered

Supplemental Fig. 6. Quantifying disorder using 'total fraction across proteome' approach. Same as Fig. 2, except 'Fraction Disorder' is quantified using 'total fraction across proteome,' which is the total number of residues predicted to be disordered in a proteome out of the total number of residues in the proteome. This contrasts with Fig. 2, which shows the average fraction of disorder per protein in a proteome. In addition, this version of the phylogenetic tree includes additional details with respect to major taxonomic lineages.
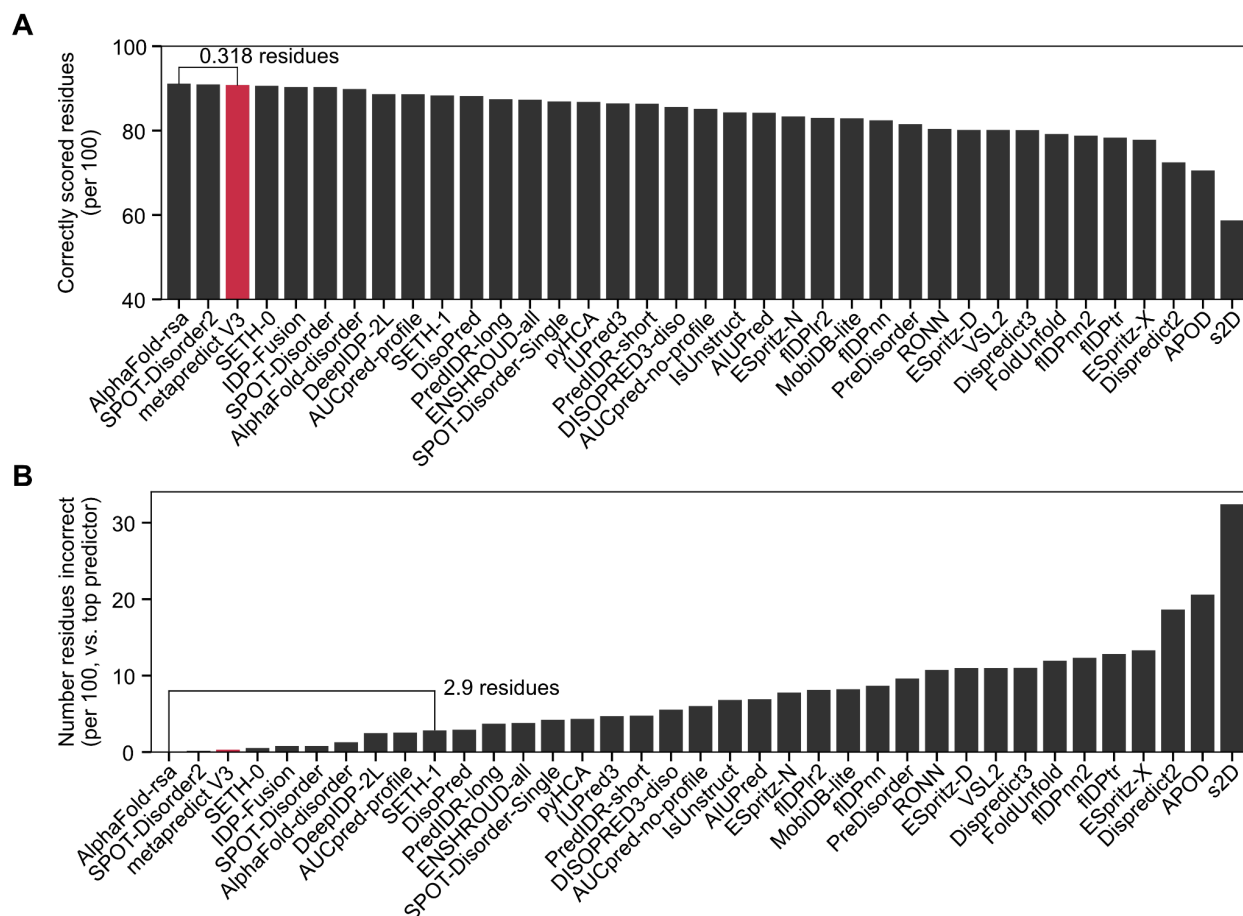
Supplemental Fig. 7. Top disorder predictors show comparable accuracy.

(A) The number of correctly scored residues per 100 residues predicted was quantified using the CAID2 Disorder-PDB analysis. Metapredict V3 is highlighted in red.

(B) Differences in the number of residues incorrectly predicted by each predictor compared to the most accurate disorder predictor quantified using the CAID2 Disorder-PDB analysis. The difference between the top disorder predictor and the tenth-best disorder predictor in terms of accuracy is shown.
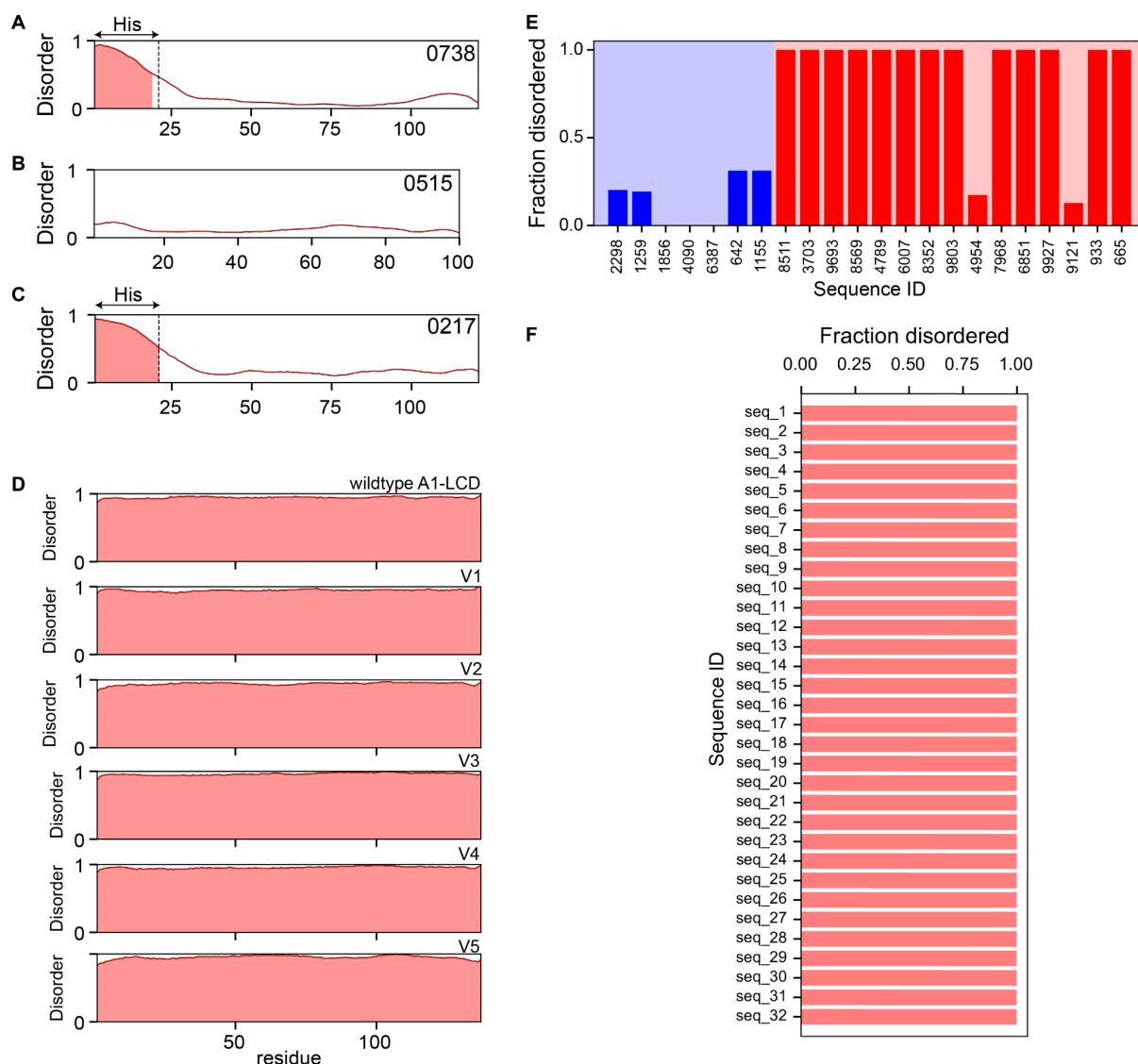
Supplemental Fig. 8. Evaluation of metapredict V3 on synthetic proteins.
(A-C) Artificial folded proteins are designed through deep hallucination[3]. In sequences (A) and (C), an N-terminal HIS tag is presented; metapredict V3 correctly predicts that the HIS tag will be disordered and the remainder of the protein will be folded.
(D) Synthetic variants of the hnRNPA1 low complexity domain (with wildtype at top) were designed by Pesce *et al.* and tested *in vitro*[4]. All variants are experimentally verified as being disordered.
(E) Random peptide sequences assessed *in vitro* by Tretyachenko *et al.*, folded (left-hand seven sequences in blue shaded region), and disordered (right-hand 15 sequences in red shaded region) were assessed[5].

(F) Thirty-two rationally designed disordered proteins were designed by Emenecker & Guadalupe *et al.*[6]. All sequences were tested in cells and found to be disordered based on inferred end-to-end distance.
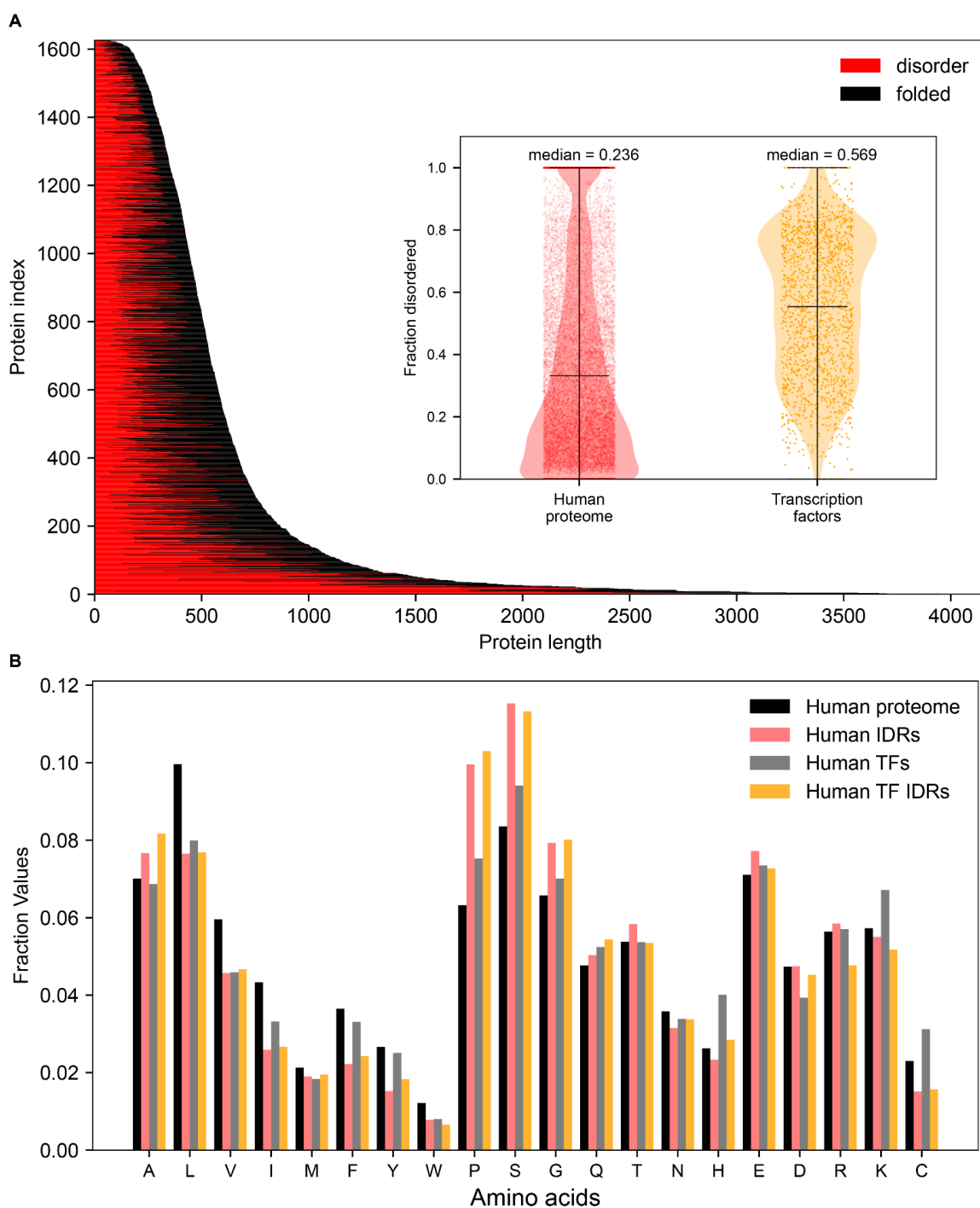
Supplemental Fig. 9. Analysis of human transcription factors.
(A) 1608 human transcription factors are represented as a stacked bar plot ordered by total protein length[7]. For each protein (indexed along the y-axis), the number of residues within IDRs is shown as red bars, while the number of residues in folded domains is black. Inset compares human transcription factors vs. human proteome; human transcriptions are (on average) ~2x

more disordered than proteins in the human proteome. Disorder in inset quantified in terms of the fraction of residues found in disordered regions.

(B) Comparison of amino acid composition in [from left-to-right] the human proteome (black), all human IDRs (red), all human transcription factors (grey), and human transcription factor IDRs. Human transcription factors see enrichment for lysine, cysteine, and histidine (found in DNA binding domains), but by and large, on average, human transcription factor IDRs and all human proteome IDRs are compositionally quite similar.

SUPPLEMENTARY REFERENCES

1. Necci, M., Piovesan, D., CAID Predictors, DisProt Curators & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* (2021). doi:10.1038/s41592-021-01117-3

2. Conte, A. D., Mehdiabadi, M., Bouhraoua, A., Miguel Monzon, A., Tosatto, S. C. E. & Piovesan, D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins* **91,** 1925–1934 (2023).

3. Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T. & Baker, D. De novo protein design by deep network hallucination. *Nature* **600,** 547–552 (2021).

4. Pesce, F., Bremer, A., Tesei, G., Hopkins, J. B., Grace, C. R., Mittag, T. & Lindorff-Larsen, K. Design of intrinsically disordered protein variants with diverse structural properties. *Sci. Adv.* **10,** eadm9926 (2024).

5. Tretyachenko, V., Vymětal, J., Bednárová, L., Kopecký, V., Jr, Hofbauerová, K., Jindrová, H., Hubálek, M., Souček, R., Konvalinka, J., Vondrášek, J. & Hlouchová, K. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **7,** 15449 (2017).

6. Emenecker, R. J., Guadalupe, K., Shamoon, N. M., Sukenik, S. & Holehouse, A. S. Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv* 2023.10.29.564547 (2023).

7. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The Human Transcription Factors. *Cell* **172,** 650–665 (2018).