

Biophysical Journal, Volume 120

Supplemental information

Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure

Ryan J. Emenecker, Daniel Griffith, and Alex S. Holehouse

Supplemental Information:

metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure

Ryan J. Emenecker^{1,2,3}, Daniel Griffith^{1,2}, Alex S. Holehouse^{1,2*}

1. Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, 63100, USA
2. Center for Science and Engineering Living Systems (CSELS), Washington University, St. Louis, MO 63130, USA
3. Center for Engineering Mechanobiology, Washington University, St. Louis, MO 63130, USA

* Correspondence: alex.holehouse@wustl.edu

Supplemental Materials and Methods

Evaluating metapredict using CheZOD scores

In line with previous work, we assessed how the continuous probability values of various predictors correlated with the CheZOD scores using the Pearson's correlation coefficient (Eq. 1) (Nielsen & Mulder, 2019). CheZOD scores increase with order and decrease with disorder. A Pearson's correlation coefficient of -1 would mean that a predictor is perfectly anti-correlated with the Z-scores, and 0 would mean that there is no correlation. As such, the results are displayed as the absolute value of the Pearson Correlation coefficient.

Based on the CheZOD, metapredict ranked 8th among the 23 predictors previously examined (**Supplemental Fig. 1A**). We also calculated the area under a receiver operating characteristic curve (AUC). The receiver operating characteristic curve uses true positive values and false positive values to assess the accuracy of the various predictors, such that a perfect predictor would have an AUC of 1. Based on this assessment metapredict was ranked 11th out of the 23 predictors evaluated (**Supplemental Fig. 1B**).

We next examined the accuracy of metapredict in predicting binary classification of either order or disorder. Previously, a CheZOD score of less than 8 was considered disordered (Nielsen & Mulder, 2019). When converting a metapredict score to binary classification, we considered any residue with a score of 0.3 or higher as disordered. For this analysis the Matthews Correlation Coefficient (MCC) was also calculated for each predictor (Eq. 2). The MCC uses a combination of false positives, false negatives, true positives, and true negatives in order to examine the accuracy of a classifier. We found that metapredict had the 8th highest MCC out of the predictors evaluated (**Supplemental Table 1**).

Metapredict implementation and usage

metapredict is written in Python 3.7+ and uses PyTorch, with the initial network trained using PARROT (Griffith & Holehouse, 2021; Paszke et al., 2019).

We designed metapredict to be as flexible and user-friendly as possible. For example, metapredict can be used as a Python library (**Supplemental Fig. 8A**), a stand-alone command-line tool (**Supplemental Fig. 8B**) or a web server (<http://metapredict.net>) (**Supplemental Fig. 8C**). Moreover, metapredict contains functionality to generate graphs or disorder scores from the command-line by directly inputting a single protein sequence, a UniProt accession number, or a FASTA file containing many sequences. Finally, in comparison to other predictors, which can take seconds, minutes or even hours per sequence, metapredict's computational performance makes it sufficiently fast that on-the-fly disorder prediction can be faster than reading pre-computed values from disk. It is this combination of accuracy, computational efficiency, ease of use, and flexibility that makes metapredict a convenient tool for any kind of disorder prediction, from single sequences to proteome-wide analyses.

To illustrate the ability of metapredict to predict consensus scores, **Supplemental Fig. 9** shows the computed consensus scores and the analogous prediction for four proteins with IDRs. Across our datasets, we found that metapredict generally performed better than over two-thirds of the currently available disorder predictors examined, likely with a slight bias for false negatives when the default disorder threshold is applied.

For a list of features see the documentation at <https://metapredict.readthedocs.io/>.

Evaluating metapredict disorder scores

Evaluations and datasets for the CheZOD score analysis (116 sequences) can be found in (Nielsen & Mulder, 2019). Evaluations and datasets for the Critical Assessment of protein Intrinsic Disorder prediction (CAID) analysis (652 sequences) can be found in (Necci et al., 2021). For convenience, all sequences and scores used are also provided at <https://github.com/holehouse-lab/supportingdata/>. All values are also found in **Supplemental tables 1-8**. Details on results including additional statistical analyses and the raw performance scores for each predictor that was used for comparisons to metapredict can be found in the supporting materials and methods.

Statistical analysis for evaluating the accuracies of disorder predictors

Statistical analysis and predictor evaluation was carried out following the protocols described previously and reproduced here for completeness (Necci et al., 2021; Nielsen & Mulder, 2019).

Predictor evaluation is performed via the Pearson's Correlation coefficient (R_p), the Matthew's Correlation Coefficient (MCC), and the F1-score.

The Pearson's Correlation Coefficient (R_p) is calculated as,

$$R_p = \frac{\sum(x_i - x_a)(y_i - y_a)}{\sqrt{\sum(x_i - x_a)^2 \sum(y_i - y_a)^2}} \quad (\text{Eq. 1})$$

where x_i is the value of the current predicted disorder value for a residue in a sequence, x_a is the mean predicted disorder value for the residues in a sequence, y_i is the actual disorder value (specifically the CheZOD score) of the current residue, and y_a is the mean actual disorder value.

The Matthew's Correlation Coefficient (MCC) is calculated as,

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FP) \times (TP + FN) \times (TN + FN)}} \quad (\text{Eq. 2})$$

Where true positives (TP) are the number of times a disorder predictor predicts a disordered residue to be disordered, true negatives (TN) are the number of times a predictor does not predict something to be disordered when it is not disordered, false positives (FP) are the number of times a predictor predicts a residue to be disordered when it is in fact not disordered, and false negatives (FN) are the number of times a predictor predicts a residue to not be disordered when it is in fact disordered.

Finally, the F1-score is calculated as

$$\text{F1-score} = \frac{TP}{TP + (0.5 \times (FP + FN))} \quad (\text{Eq. 3})$$

Where TP, FP, and FN are defined above.

Metapredict training of disorder predictor

The metapredict disorder prediction was trained using PARROT with slight modifications to the default settings (Griffith & Holehouse, 2021). We set the number of training epochs, which defines the number of times the complete dataset is assessed and used to update the network parameters, to 100. Increasing the number of epochs further gave a negligible improvement in performance. The PARROT data type was set to 'residues' and the number of classes was set to '1' (for regression). The learning rate (--learning-rate), which is a parameter that alters the rate that the model updates weights after each round of back-propagation, was set to 0.001. The number of layers (--num-layers), which is the number of layers in the network between the input layer and the output layer, was set to 1. The hidden vector size (--hidden-size), which is the size of hidden vectors within the BRNN, was set to 5. The batch size (--batch), which is the number of sequences processed at the same time, was set to 32. For training, validation, and testing, 70% of the data was used for training, 15% of the data was used for validation, and 15% of the data was used for testing.

The proteomes for which consensus disorder scores were available at the time of training were: *Danio rerio* (UP000000437, 43,841 proteins), *Gallus gallus* (UP000000539, 25,238), *Mus*

musculus (UP000000589, 44,470 proteins), *Drosophila melanogaster* (UP000000803, 21,114 proteins), *Dictyostelium discoideum* (UP000002195, 12,733 proteins), *Canis lupus familiaris* (UP000002254, 45,089 proteins), *Saccharomyces cerevisiae* (UP000002311, 6,049 proteins), *Rattus norvegicus* (UP000002494, 29,090 proteins), *Homo sapiens* (UP000005640, 66,835), *Arabidopsis thaliana* (UP000006548, 39,342 proteins), *Sus scrofa* (UP000008227, 49,792 proteins), and *Bos taurus* (UP000009136, 37,367 proteins). These numbers reflect protein sequences composed of the 20 standard amino acids only. Cross-referencing the training dataset (70% of the total sequences) taken from these proteomes against the assessment databases used (CheZOD and CAID) identified 28/116 from CheZOD and 451/652 from CAID databases in a total training set of ~295,000 sequences. These proteomes and the associated consensus disorder scores were obtained from MobiDB, but were originally curated by UniProt (Acids Research & 2021, 2021; Piovesan et al., 2021; UniProt Consortium, 2019).

Metapredict performance

Metapredict has no specific hardware requirements and performs well across all set ups tested. Hardware tested included an Ubuntu-running Dell desktop (Ubuntu 18.04 with Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz CPU, 32 GBs DDR4 RAM, with a Toshiba 512 GB SSD (KBG40ZNS512G), a 2020 Apple Mac Mini (16 GB unified memory, Apple M1 processor), a 2019 16-inch MacBook Pro (64 GB 2667 MHz DDR4 RAM, 2.3 GHz Intel Core i9 processor, Intel UHD Graphics 630 integrated graphics), and a 2012 MacBook Pro (2.9 GHz Intel Core i7 processor, 8 GB 1600 MHz DDR3 RAM, Intel HD Graphics 4000 1536 MB integrated graphics). Importantly, as evident by our testing on a 2012 MacBook Pro (which scored at 7,238 residues per second), metapredict does not require a high-end modern computer to be fast. Even on the most basic virtual machine available, (Ubuntu 18.04 with single virtual Intel CPU (2.4 GHz), 1 GB DIMM memory, 20 GB SSD), metapredict performs at ~6000 residues per second.

To compare AUCPreD vs. metapredict (**Supplemental Fig. S1**) predictions were performed on a desktop machine running Ubuntu 18.04 with an Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz CPU, 32 GBs DDR4 RAM, with a Toshiba 512 GB SSD (KBG40ZNS512G). To ensure a fair comparison, each sequence was isolated and placed in its own FASTA file, and the predictor ran on each file independently. In reality, if one had multiple sequences, placing them into a single FASTA file would be a more efficient approach to minimize the amount of time reading from the filesystem. While the effective contribution of file read time is negligible for AUCPreD, in all cases for metapredict it is the dominant determinant of execution time.

To compare metapredict execution time against other predictors based on times reported in the CAID experiment, we used per-sequence execution times for AUCPreD on our local hardware and on the CAID hardware to calibrate and approximate conversion factor. A more detailed explanation of this process is described in https://github.com/holehouse-lab/supportingdata/tree/master/2021/emenecker_metapredict_2021/performance/metapredict_on_caid_disprot

AlphaFold2 pLDDT predictor data

We obtained pLDDT scores for every AlphaFold2 model from <http://ftp.ebi.ac.uk/pub/databases/alphafold/> and extracted the pLDDT score from these models for twenty one proteomes (UP000002485_284812_SCHPO, UP000000805_243232_METJA, UP000001450_36329_PLAF7, UP000005640_9606_HUMAN, UP000001584_83332_MYCTU, UP000001940_6239_CAEEL, UP000000625_83333_ECOLI, UP000002296_353153_TRYCC, UP000000803_7227_DROME, UP000007305_4577_MAIZE, UP000002195_44689_DICDI, UP000002311_559292_YEAST, UP000002494_10116_RAT, UP000008153_5671_LEIIN, UP000008816_93061_STAA8, UP000006548_3702_ARATH, UP000008827_3847_SOYBN, UP000000437_7955_DANRE, UP000000589_10090_MOUSE, UP000059680_39947_ORYSJ, UP000000559_237561_CANAL). For each of these proteomes all PDB models were used, with the exception of the human proteome where only the first (F1) models were used. In total 363,264 different sequences were used.

AlphaFold2 pLDDT predictor training

The AlphaFold2 (AF2) pLDDT predictor (alphaPredict) was trained using PARROT (version 1.5.0), the same general-purpose deep learning framework used to create the BRNN behind the metapredict disorder predictions. The parameters used for PARROT training were as follows: residues datatype, 1 class (for regression), a learning-rate of 0.001, 2 hidden layers, 20 hidden vectors, a batch size of 32, and 200 training epochs.

AlphaFold2 pLDDT predictor accuracy

At the time of writing, our AF2 pLDDT confidence score predictor obtained an r^2 value for the actual versus predicted scores in the test set was 0.7148 and the average error per residue was approximately 9.20% (network V4).

AlphaFold2 pLDDT predictor implementation

For improved modularity, the AlphaFold2 predictor is currently implemented as a separate Python package (alphaPredict) which is encoded as a silent dependency to metapredict (<https://github.com/ryanemenecker/alphaPredict>). alphaPredict can also be downloaded and used independently from metapredict in its own right. alphaPredict is written in Python 3.7+ and uses PyTorch, with the initial network trained using PARROT (Griffith & Holehouse, 2021; Paszke et al., 2019).

AlphaFold2 pLDDT predictor improvements

The prior information regarding the network (network V2) used for the AF2 pLDDT confidence score predictions in metapredict is up to date at the time of this writing. However, additional networks (a V3 and V4) are currently in progress. However, due to the increased number of proteomes used for training these networks and the slightly altered parameters, they currently have weeks of additional training remaining before completion. For the most up to date information on the AF2 pLDDT prediction in metapredict, please see <https://github.com/ryanemenecker/alphaPredict>.

Supplemental Tables

	Rp	AUC	MCC	TPR	TNR	FPR	FNR	prec	Acc
SPOT-dis	-0.657	0.881	0.59089	0.7067	0.8757	0.1243	0.2933	0.9177	0.7912
MFDp2	-0.631	0.853	0.578	0.8528	0.7275	0.2725	0.1472	0.8599	0.7901
MetaDisorder	-0.617	0.865	0.541	0.8221	0.7287	0.2713	0.1779	0.856	0.7754
AUCpreD	-0.598	0.865	0.54	0.7154	0.8555	0.1445	0.2846	0.9066	0.7854
MetaDisorderMD2	-0.614	0.852	0.54	0.7688	0.7955	0.2045	0.2312	0.8806	0.7821
MetaDisorderMD	-0.616	0.853	0.519	0.7246	0.8227	0.1773	0.2754	0.8891	0.7736
IUPred_long	-0.566	0.834	0.512	0.7477	0.7887	0.2113	0.2523	0.8741	0.7682
AUCpreD_noEvo	-0.512	0.841	0.48658	0.603	0.8664	0.1336	0.397	0.8985	0.7347
metapredict	-0.559	0.832	0.489	0.790	0.718	0.282	0.210	0.860	0.754
DISPROT (VSL2b)	-0.536	0.808	0.466	0.8525	0.5986	0.4014	0.1475	0.8064	0.7256
RONN	-0.5	0.804	0.455	0.7387	0.7368	0.2632	0.2613	0.8463	0.7377
DISOPRED3	-0.551	0.833	0.451	0.5301	0.9207	0.0793	0.4699	0.9292	0.7254
IUPred_short	-0.532	0.822	0.429	0.6284	0.8249	0.1751	0.3716	0.8756	0.7267
ESpritz_DisProt	-0.419	0.748	0.412	0.8208	0.5846	0.4154	0.1792	0.7949	0.7027
PrDOS	-0.541	0.836	0.409	0.5333	0.889	0.111	0.4667	0.9041	0.7111
ESpritz_NMR	-0.478	0.797	0.374	0.4662	0.9087	0.0913	0.5338	0.9092	0.6874
DISpro	-0.437	0.805	0.35712	0.3185	0.9565	0.0435	0.6815	0.9349	0.6375
DISOPRED2	-0.33	0.738	0.3419	0.642	0.6994	0.3006	0.358	0.8073	0.6707
MetaDisorder3D	-0.361	0.727	0.333	0.5841	0.7674	0.2326	0.4159	0.8312	0.6757
DisEMBL_coils	-0.404	0.735	0.32932	0.7252	0.6016	0.3984	0.2748	0.7812	0.6634
ESpritz_Xray	-0.438	0.791	0.286	0.282	0.9597	0.0403	0.718	0.9321	0.6208
DisEMBL_remark465	-0.386	0.737	0.284	0.4219	0.8598	0.1402	0.5781	0.8551	0.6409
DisEMBL_hotloops	-0.286	0.702	0.239	0.432	0.8093	0.1907	0.568	0.8163	0.6206

Supplemental Table 1. CHEZOD scores obtained from (Nielsen & Mulder, 2019).

Rankings

Rank	<u>Rp</u>	<u>AUC</u>	<u>MCC</u>	<u>prec</u>	<u>Acc</u>
1	SPOT-dis	SPOT-dis	SPOT-dis	DISpro	SPOT-dis
2	MFDp2	AUCpreD	MFDp2	ESpritz_Xray	MFDp2
3	MetaDisorder	MetaDisorder	MetaDisorder	DISOPRED3	AUCpreD
4	MetaDisorderMD	MFDp2	AUCpreD	SPOT-dis	MetaDisorderMD2
5	MetaDisorderMD2	MetaDisorderMD	MetaDisorderMD2	ESpritz_NMR	MetaDisorder
6	AUCpreD	MetaDisorderMD2	MetaDisorderMD	AUCpreD	MetaDisorderMD
7	IUPred_long	AUCpreD_noEvo	IUPred_long	PrDOS	IUPred_long
8	metapredict	PrDOS	metapredict	AUCpreD_noEvo	metapredict
9	DISOPRED3	IUPred_long	AUCpreD_noEvo	MetaDisorderMD	RONN
10	PrDOS	DISOPRED3	DISPROT (VSL2b)	MetaDisorderMD2	AUCpreD_noEvo
11	DISPROT (VSL2b)	metapredict	RONN	IUPred_short	IUPred_short
12	IUPred_short	IUPred_short	DISOPRED3	IUPred_long	DISPROT (VSL2b)
13	AUCpreD_noEvo	DISPROT (VSL2b)	IUPred_short	MFDp2	DISOPRED3
14	RONN	DISpro	ESpritz_DisProt	metapredict	PrDOS
15	ESpritz_NMR	RONN	PrDOS	MetaDisorder	ESpritz_DisProt
16	ESpritz_Xray	ESpritz_NMR	ESpritz_NMR	DisEMBL_remark465	ESpritz_NMR
17	DISpro	ESpritz_Xray	DISpro	RONN	MetaDisorder3D
18	ESpritz_DisProt	ESpritz_DisProt	DISOPRED2	MetaDisorder3D	DISOPRED2
19	DisEMBL_coils	DISOPRED2	MetaDisorder3D	DisEMBL_hotloops	DisEMBL_coils
20	DisEMBL_remark465	DisEMBL_remark465	DisEMBL_coils	DISOPRED2	DisEMBL_remark465
21	MetaDisorder3D	DisEMBL_coils	ESpritz_Xray	DISPROT (VSL2b)	DISpro
22	DISOPRED2	MetaDisorder3D	DisEMBL_remark465	ESpritz_DisProt	ESpritz_Xray
23	DisEMBL_hotloops	DisEMBL_hotloops	DisEMBL_hotloops	DisEMBL_coils	DisEMBL_hotloops

Supplemental Table 2. CHEZOD-based rankings. This table is provided in a .xlsx format at the manuscript's GitHub repository.

	TN	FP	FN	TP	MCC	F1-s	TNR	TPR	PPV	BAC
fIDPnn	585	16	19	26	0.569	0.598	0.973	0.578	0.619	0.776
RawMSA	582	19	19	26	0.546	0.578	0.968	0.578	0.578	0.773
VSL2B	578	23	22	23	0.468	0.505	0.962	0.511	0.5	0.736
fIDPIr	566	35	18	27	0.468	0.505	0.942	0.6	0.435	0.771
Predisorder	589	12	26	19	0.479	0.5	0.98	0.422	0.613	0.701
SPOT-Disorder1	572	29	23	22	0.416	0.458	0.952	0.489	0.431	0.72
DisoMine	551	50	17	28	0.421	0.455	0.917	0.622	0.359	0.77
AUCpreD	588	13	28	17	0.431	0.453	0.978	0.378	0.567	0.678
SPOT-Disorder2	574	27	24	21	0.409	0.452	0.955	0.467	0.438	0.711
metapredict	599	8	25	20	0.539	0.548	0.987	0.444	0.714	0.716
SPOT-Disorder-Single	594	7	30	15	0.452	0.448	0.988	0.333	0.682	0.661
IsUnstruct	588	13	29	16	0.411	0.432	0.978	0.356	0.552	0.667
IUPred2A-long	595	6	32	13	0.42	0.406	0.99	0.289	0.684	0.639
Gene3D	505	96	10	35	0.391	0.398	0.84	0.778	0.267	0.809
ESpritz-N	597	4	33	12	0.426	0.393	0.993	0.267	0.75	0.63
ESpritz-D	555	46	23	22	0.342	0.389	0.923	0.489	0.324	0.706
PyHCA	596	5	33	12	0.411	0.387	0.992	0.267	0.706	0.629
JRONN	595	6	33	12	0.397	0.381	0.99	0.267	0.667	0.628
MobiDB-lite	599	2	34	11	0.437	0.379	0.997	0.244	0.846	0.621
DisPredict-2	586	15	32	13	0.33	0.356	0.975	0.289	0.464	0.632
IUPred2A-short	599	2	35	10	0.413	0.351	0.997	0.222	0.833	0.609
S2D-2	572	29	30	15	0.288	0.337	0.952	0.333	0.341	0.643
PDB(observed)	468	133	13	32	0.286	0.305	0.779	0.711	0.194	0.745
AUCpreD-np	590	11	35	10	0.293	0.303	0.982	0.222	0.476	0.602
ESpritz-X	595	6	36	9	0.321	0.3	0.99	0.2	0.6	0.595
FoldUnfold	456	145	14	31	0.256	0.281	0.759	0.689	0.176	0.724
DISOPRED-3.1	596	5	39	6	0.246	0.214	0.992	0.133	0.545	0.563

DisEMBL-HL	601	0	41	4	0.288	0.163	1	0.089	1	0.544
PDB(remote)	590	11	42	3	0.085	0.102	0.982	0.067	0.214	0.524
DisEMBL-465	601	0	43	2	0.204	0.085	1	0.044	1	0.522
PDB(close)	589	12	43	2	0.043	0.068	0.98	0.044	0.143	0.512
Conservation	441	160	38	7	-0.064	0.066	0.734	0.156	0.042	0.445
DynaMine	601	0	45	0	0	0	1	0	0	0.5
GlobPlot	601	0	45	0	0	0	1	0	0	0.5
DFLpred	601	0	45	0	0	0	1	0	0	0.5

Supplemental Table 3. CAID scores taken from (Necci et al., 2021). This table is provided in a .xlsx format at the manuscript's GitHub repository.

Rankings

Rank	MCC	F1	BAC	PPV
1	fIDPnn	fIDPnn	Gene3D	DisEMBL-HL
2	RawMSA	RawMSA	fIDPnn	DisEMBL-465
3	metapredict	metapredict	RawMSA	MobiDB-lite
4	Predisorder	VSL2B	fIDPIr	IUPred2A-short
5	VSL2B	fIDPIr	DisoMine	ESpritz-N
6	fIDPIr	Predisorder	PDB(observed)	metapredict
7	SPOT-Disorder-Single	SPOT-Disorder1	VSL2B	PyHCA
8	MobiDB-lite	DisoMine	FoldUnfold	IUPred2A-long
9	AUCpreD	AUCpreD	SPOT-Disorder1	SPOT-Disorder-Single
10	ESpritz-N	SPOT-Disorder2	metapredict	JRONN
11	DisoMine	SPOT-Disorder-Single	SPOT-Disorder2	fIDPnn
12	IUPred2A-long	IsUnstruct	ESpritz-D	Predisorder
13	SPOT-Disorder1	IUPred2A-long	Predisorder	ESpritz-X
14	IUPred2A-short	Gene3D	AUCpreD	RawMSA
15	IsUnstruct	ESpritz-N	IsUnstruct	AUCpreD
16	PyHCA	ESpritz-D	SPOT-Disorder-Single	IsUnstruct
17	SPOT-Disorder2	PyHCA	S2D-2	DISOPRED-3.1
18	JRONN	JRONN	IUPred2A-long	VSL2B
19	Gene3D	MobiDB-lite	DisPredict-2	AUCpreD-np
20	ESpritz-D	DisPredict-2	ESpritz-N	DisPredict-2
21	DisPredict-2	IUPred2A-short	PyHCA	SPOT-Disorder2
22	ESpritz-X	S2D-2	JRONN	fIDPIr
23	AUCpreD-np	PDB(observed)	MobiDB-lite	SPOT-Disorder1
24	S2D-2	AUCpreD-np	IUPred2A-short	DisoMine
25	DisEMBL-HL	ESpritz-X	AUCpreD-np	S2D-2
26	PDB(observed)	FoldUnfold	ESpritz-X	ESpritz-D
27	FoldUnfold	DISOPRED-3.1	DISOPRED-3.1	Gene3D

28	DISOPRED-3.1	DisEMBL-HL	DisEMBL-HL	PDB(remote)
29	DisEMBL-465	PDB(remote)	PDB(remote)	PDB(observed)
30	PDB(remote)	DisEMBL-465	DisEMBL-465	FoldUnfold
31	PDB(close)	PDB(close)	PDB(close)	PDB(close)
32	DynaMine	Conservation	DynaMine	Conservation
33	GlobPlot	DynaMine	GlobPlot	DynaMine
34	DFLpred	GlobPlot	DFLpred	GlobPlot
35	Conservation	DFLpred	Conservation	DFLpred

Supplemental Table 4. CAID-based rankings. This table is provided in a .xlsx format at the manuscript's GitHub repository.

	BAC	F1-S	FPR	MCC	PPV	TPR	TNR
fIDPnn	0.72	0.483	0.189	0.37	0.392	0.629	0.811
SPOT-Disorder2	0.725	0.469	0.343	0.349	0.333	0.794	0.657
fIDPIr	0.693	0.452	0.184	0.33	0.374	0.57	0.816
RawMSA	0.714	0.445	0.297	0.328	0.321	0.725	0.703
Predisorder	0.691	0.435	0.28	0.301	0.324	0.661	0.72
SPOT-Disorder1	0.723	0.434	0.386	0.33	0.294	0.832	0.614
AUCpreD	0.712	0.433	0.376	0.318	0.297	0.801	0.624
SPOT-Disorder-Single	0.71	0.432	0.341	0.315	0.302	0.76	0.659
ESpritz-D	0.703	0.428	0.325	0.307	0.303	0.731	0.675
AUCpreD-np	0.699	0.424	0.327	0.301	0.3	0.725	0.673
DisoMine	0.698	0.424	0.326	0.299	0.3	0.721	0.674
metapredict	0.693	0.421	0.320	0.293	0.300	0.705	0.680
IUPred2A-short	0.688	0.42	0.297	0.29	0.305	0.674	0.703
MobiDB-lite	0.688	0.42	0.296	0.289	0.305	0.673	0.704
IUPred-long	0.686	0.418	0.294	0.287	0.305	0.666	0.706
IUPred-short	0.688	0.418	0.304	0.288	0.302	0.679	0.696
ESpritz-X	0.689	0.418	0.309	0.288	0.301	0.686	0.691
IsUnstruct	0.689	0.418	0.311	0.288	0.3	0.688	0.689
IUPred2A-long	0.685	0.416	0.299	0.285	0.302	0.67	0.701
VSL2B	0.684	0.408	0.341	0.277	0.286	0.709	0.659
JRONN	0.672	0.401	0.318	0.263	0.287	0.663	0.682
ESpritz-N	0.664	0.4	0.271	0.259	0.3	0.599	0.729
DISOPRED-3.1	0.674	0.393	0.401	0.258	0.266	0.749	0.599
PyHCA	0.66	0.385	0.346	0.241	0.271	0.666	0.654
DynaMine	0.66	0.384	0.362	0.24	0.267	0.682	0.638
Gene3D	0.653	0.368	0.486	0.226	0.24	0.791	0.514
FoldUnfold	0.642	0.365	0.382	0.211	0.251	0.666	0.618
DisEMBL-465	0.627	0.363	0.215	0.214	0.296	0.468	0.785

S2D-1	0.633	0.361	0.329	0.203	0.259	0.595	0.671
PDB (close)	0.637	0.353	0.38	0.202	0.242	0.655	0.62
S2D-2	0.624	0.347	0.439	0.183	0.232	0.687	0.561
PDB (observed)	0.616	0.339	0.565	0.174	0.215	0.796	0.435
DisPredict-2	0.599	0.326	0.326	0.152	0.237	0.523	0.674
PDB (remote)	0.614	0.321	0.45	0.163	0.21	0.678	0.55
GlobPlot	0.587	0.312	0.253	0.143	0.246	0.427	0.747
Conservation	0.552	0.288	0.483	0.077	0.191	0.587	0.517
DisEMBL-HL	0.577	0.286	0.099	0.172	0.33	0.253	0.901
DFLpred	0.503	0.025	0.008	0.022	0.249	0.013	0.992

Supplemental Table 5. CAID per-residue assessment using DisProt dataset. This table is provided in a .xlsx format at the manuscript's GitHub repository.

Rankings

Rank	MCC	F1	BAC	PPV
1	fIDPnn	fIDPnn	SPOT-Disorder2	fIDPnn
2	SPOT-Disorder2	SPOT-Disorder2	SPOT-Disorder1	fIDPIr
3	SPOT-Disorder1	fIDPIr	fIDPnn	SPOT-Disorder2
4	fIDPIr	RawMSA	RawMSA	DisEMBL-HL
5	RawMSA	Predisorder	AUCpreD	Predisorder
6	AUCpreD	SPOT-Disorder1	SPOT-Disorder-Single	RawMSA
7	SPOT-Disorder-Single	AUCpreD	ESpritz-D	IUPred2A-short
8	ESpritz-D	SPOT-Disorder-Single	AUCpreD-np	MobiDB-lite
9	AUCpreD-np	ESpritz-D	DisoMine	IUPred-long
10	Predisorder	AUCpreD-np	fIDPIr	ESpritz-D
11	DisoMine	DisoMine	metapredict	SPOT-Disorder-Single
12	metapredict	metapredict	Predisorder	IUPred-short
13	IUPred2A-short	IUPred2A-short	ESpritz-X	IUPred2A-long
14	MobiDB-lite	MobiDB-lite	IsUnstruct	ESpritz-X
15	ESpritz-X	ESpritz-X	IUPred2A-short	AUCpreD-np
16	IsUnstruct	IsUnstruct	MobiDB-lite	DisoMine
17	IUPred-short	IUPred-short	IUPred-short	IsUnstruct
18	IUPred-long	IUPred-long	IUPred-long	ESpritz-N
19	IUPred2A-long	IUPred2A-long	IUPred2A-long	metapredict
20	VSL2B	VSL2B	VSL2B	AUCpreD
21	JRONN	JRONN	DISOPRED-3.1	DisEMBL-465
22	ESpritz-N	ESpritz-N	JRONN	SPOT-Disorder1
23	DISOPRED-3.1	DISOPRED-3.1	ESpritz-N	JRONN
24	PyHCA	PyHCA	PyHCA	VSL2B
25	DynaMine	DynaMine	DynaMine	PyHCA
26	Gene3D	Gene3D	Gene3D	DynaMine

27	DisEMBL-465	FoldUnfold	FoldUnfold	DISOPRED-3.1
28	FoldUnfold	DisEMBL-465	PDB (close)	S2D-1
29	S2D-1	S2D-1	S2D-1	FoldUnfold
30	PDB (close)	PDB (close)	DisEMBL-465	DFLpred
31	S2D-2	S2D-2	S2D-2	GlobPlot
32	PDB (observed)	PDB (observed)	PDB (observed)	PDB (close)
33	DisEMBL-HL	DisPredict-2	PDB (remote)	Gene3D
34	PDB (remote)	PDB (remote)	DisPredict-2	DisPredict-2
35	DisPredict-2	GlobPlot	GlobPlot	S2D-2
36	GlobPlot	Conservation	DisEMBL-HL	PDB (observed)
37	Conservation	DisEMBL-HL	Conservation	PDB (remote)
38	DFLpred	DFLpred	DFLpred	Conservation

Supplemental Table 6. CAID per-residue rankings using DisProt dataset. This table is provided in a .xlsx format at the manuscript's GitHub repository.

	BAC	F1-S	FPR	MCC	PPV	TPR	TNR
PDB(observed)	0.898	0.886	0	0.854	1	0.796	1
SPOT-Disorder1	0.846	0.788	0.09	0.696	0.795	0.782	0.91
SPOT-Disorder2	0.836	0.784	0.055	0.706	0.851	0.727	0.945
AUCpreD	0.816	0.756	0.07	0.662	0.82	0.701	0.93
PDB (close)	0.811	0.755	0.033	0.689	0.891	0.655	0.967
SPOT-Disorder-Single	0.817	0.751	0.095	0.646	0.775	0.729	0.905
RawMSA	0.815	0.745	0.106	0.635	0.755	0.736	0.894
AUCpreD-np	0.797	0.725	0.092	0.615	0.769	0.686	0.908
DISOPRED-3.1	0.796	0.724	0.092	0.613	0.768	0.684	0.908
Predisorder	0.788	0.717	0.067	0.619	0.813	0.642	0.933
IUPred-long	0.783	0.704	0.096	0.588	0.754	0.661	0.904
fIDPnn	0.782	0.701	0.113	0.576	0.727	0.676	0.887
IsUnstruct	0.779	0.7	0.091	0.585	0.76	0.648	0.909
IUPred2A-long	0.776	0.697	0.087	0.584	0.766	0.64	0.913
VSL2B	0.774	0.695	0.087	0.581	0.765	0.636	0.913
ESpritz-X	0.778	0.695	0.119	0.566	0.717	0.675	0.881
IUPred-short	0.775	0.693	0.104	0.571	0.738	0.654	0.896
DisoMine	0.78	0.693	0.16	0.55	0.668	0.721	0.84
Gene3D	0.785	0.692	0.22	0.539	0.615	0.791	0.78
IUPred2A-short	0.773	0.691	0.094	0.574	0.752	0.64	0.906
ESpritz-D	0.778	0.69	0.166	0.544	0.66	0.723	0.834
metapredict	0.791	0.712	0.124	0.585	0.718	0.705	0.876
MobiDB-lite	0.764	0.683	0.063	0.583	0.806	0.592	0.937
fIDPIr	0.761	0.671	0.119	0.537	0.705	0.641	0.881
ESpritz-N	0.751	0.662	0.073	0.554	0.779	0.575	0.927
JRONN	0.751	0.661	0.081	0.546	0.762	0.583	0.919
DynaMine	0.739	0.641	0.11	0.505	0.704	0.588	0.89
FoldUnfold	0.736	0.636	0.193	0.462	0.608	0.666	0.807

PyHCA	0.731	0.629	0.107	0.494	0.704	0.569	0.893
S2D-1	0.724	0.617	0.089	0.494	0.728	0.536	0.911
S2D-2	0.703	0.591	0.253	0.386	0.536	0.658	0.747
PDB (remote)	0.703	0.579	0.273	0.377	0.505	0.678	0.727
DisEMBL-465	0.694	0.57	0.11	0.426	0.667	0.498	0.89
DisEMBL-HL	0.641	0.535	0.47	0.262	0.415	0.752	0.53
DisPredict-2	0.625	0.491	0.285	0.24	0.455	0.534	0.715
Conservation	0.618	0.485	0.296	0.227	0.445	0.533	0.704
GlobPlot	0.641	0.48	0.111	0.328	0.613	0.394	0.889
DFLpred	0.504	0.027	0.005	0.043	0.53	0.014	0.995

Supplemental Table 7. CAID per-residue rankings using DisProt-PDB dataset. This table is provided in a .xlsx format at the manuscript's GitHub repository.

Rankings

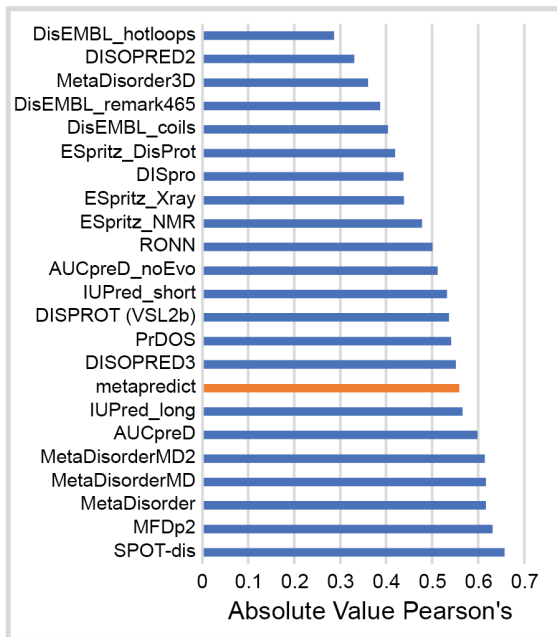
Rank	MCC	F1	BAC	PPV
1	PDB(observed)	PDB(observed)	PDB(observed)	PDB(observed)
2	SPOT-Disorder2	SPOT-Disorder1	SPOT-Disorder1	PDB (close)
3	SPOT-Disorder1	SPOT-Disorder2	SPOT-Disorder2	SPOT-Disorder2
4	PDB (close)	AUCpreD	SPOT-Disorder-Single	AUCpreD
5	AUCpreD	PDB (close)	AUCpreD	Predisorder
6	SPOT-Disorder-Single	SPOT-Disorder-Single	RawMSA	MobiDB-lite
7	RawMSA	RawMSA	PDB (close)	SPOT-Disorder1
8	Predisorder	AUCpreD-np	AUCpreD-np	ESpritz-N
9	AUCpreD-np	DISOPRED-3.1	DISOPRED-3.1	SPOT-Disorder-Single
10	DISOPRED-3.1	Predisorder	metapredict	AUCpreD-np
11	IUPred-long	metapredict	Predisorder	DISOPRED-3.1
12	IsUnstruct	IUPred-long	Gene3D	IUPred2A-long
13	metapredict	fIDPnn	IUPred-long	VSL2B
14	IUPred2A-long	IsUnstruct	fIDPnn	JRONN
15	MobiDB-lite	IUPred2A-long	DisoMine	IsUnstruct
16	VSL2B	VSL2B	IsUnstruct	RawMSA
17	fIDPnn	ESpritz-X	ESpritz-X	IUPred-long
18	IUPred2A-short	IUPred-short	ESpritz-D	IUPred2A-short
19	IUPred-short	DisoMine	IUPred2A-long	IUPred-short
20	ESpritz-X	Gene3D	IUPred-short	S2D-1
21	ESpritz-N	IUPred2A-short	VSL2B	fIDPnn
22	DisoMine	ESpritz-D	IUPred2A-short	metapredict
23	JRONN	MobiDB-lite	MobiDB-lite	ESpritz-X
24	ESpritz-D	fIDPIr	fIDPIr	fIDPIr
25	Gene3D	ESpritz-N	ESpritz-N	DynaMine
26	fIDPIr	JRONN	JRONN	PyHCA
27	DynaMine	DynaMine	DynaMine	DisoMine

28	PyHCA	FoldUnfold	FoldUnfold	DisEMBL-465
29	S2D-1	PyHCA	PyHCA	ESpritz-D
30	FoldUnfold	S2D-1	S2D-1	Gene3D
31	DisEMBL-465	S2D-2	S2D-2	GlobPlot
32	S2D-2	PDB (remote)	PDB (remote)	FoldUnfold
33	PDB (remote)	DisEMBL-465	DisEMBL-465	S2D-2
34	GlobPlot	DisEMBL-HL	DisEMBL-HL	DFLpred
35	DisEMBL-HL	DisPredict-2	GlobPlot	PDB (remote)
36	DisPredict-2	Conservation	DisPredict-2	DisPredict-2
37	Conservation	GlobPlot	Conservation	Conservation
38	DFLpred	DFLpred	DFLpred	DisEMBL-HL

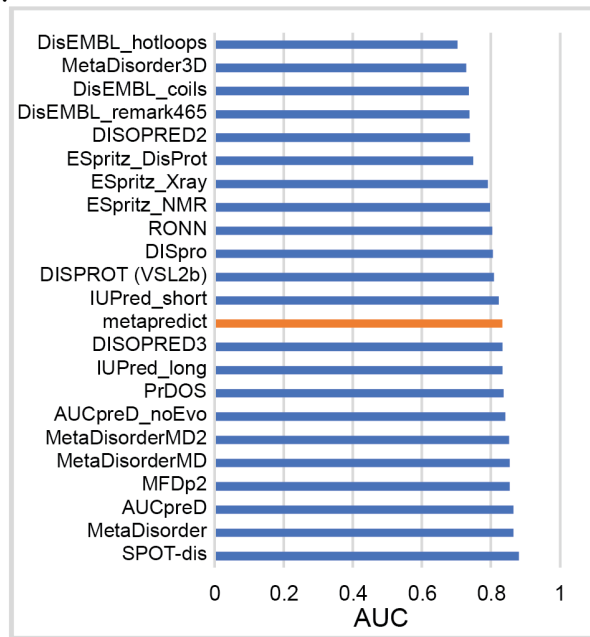
Supplemental Table 8. CAID per-residue rankings using DisProt-PDB dataset. This table is provided in a .xlsx format at the manuscript's GitHub repository.

Supplemental Figures

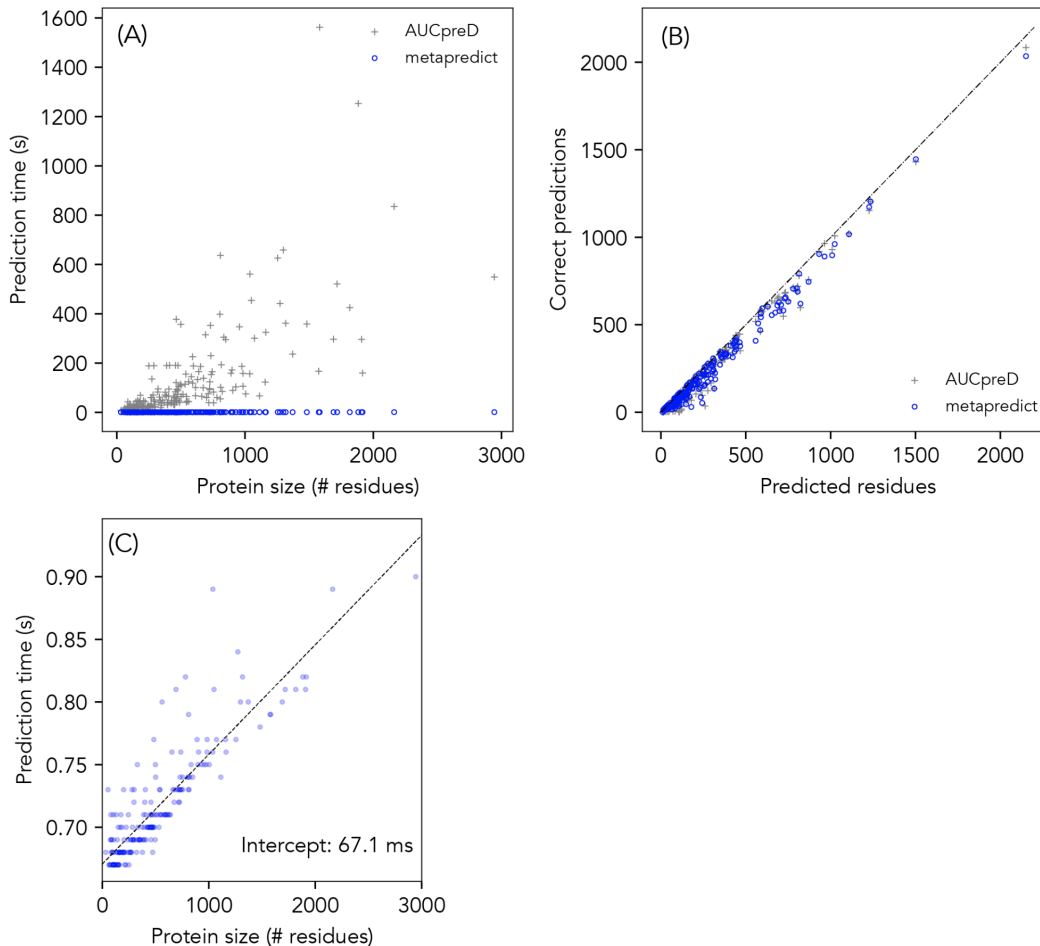
A.



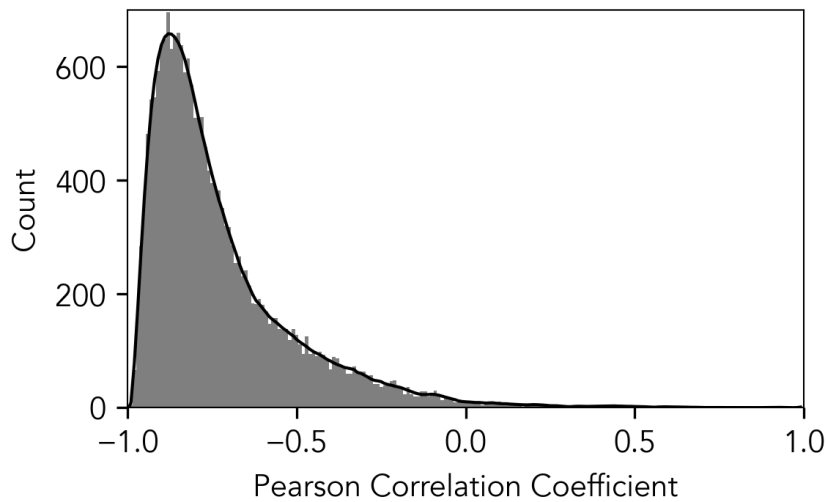
B.



Supplemental Figure 1. Evaluation of metapredict using CheZOD scores. (A) The absolute value of the Pearson's correlation coefficient calculated by comparing the correlation between each predictor's score per residue and the CheZOD score. **(B)** The area under the receiver operating characteristic curve (AUC) (generated by comparing disorder scores of various predictors to disorder predictions from CheZOD scores. Values for all predictors in (A) and (B) other than metapredict (orange bar) were obtained from (Nielsen & Mulder, 2019).

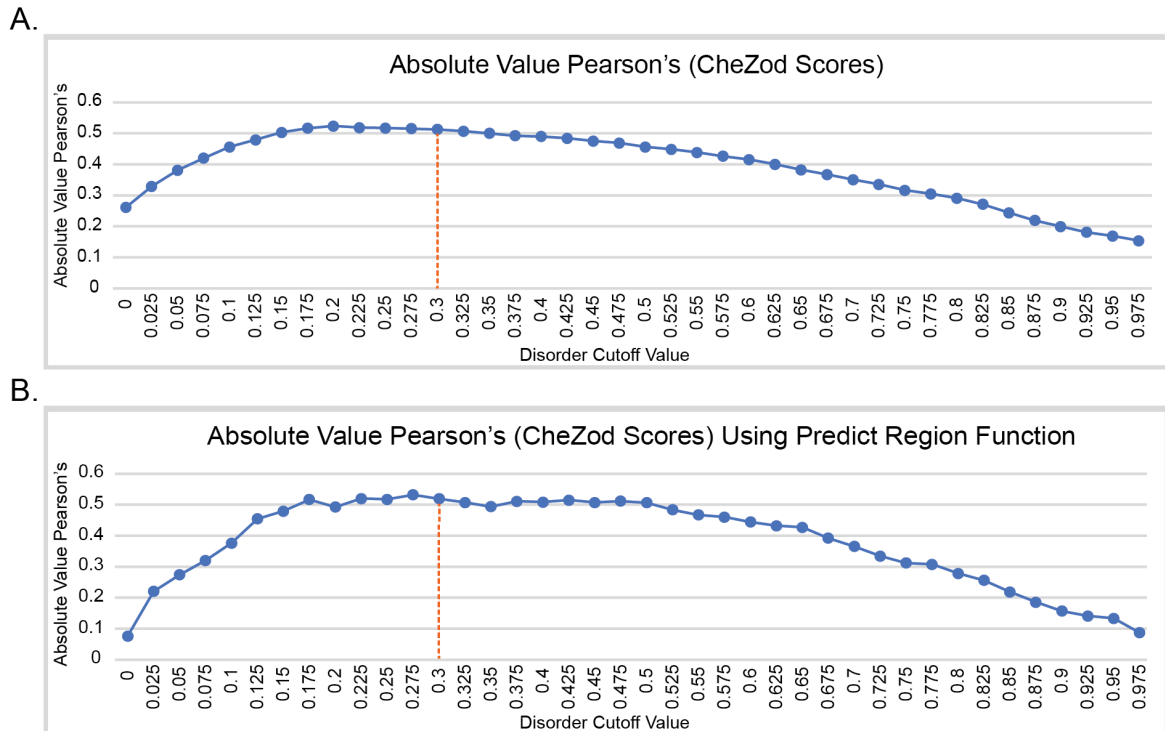


Supplemental Figure 2. Comparison of execution time and predict performance for AUCPreD vs. metapredict (A) Run time of 200 different proteins from the DISPROT dataset spanning a variety of different sizes as assessed by AUCPreD (grey crosses) vs. metapredict (blue circles). Both methods show a clear correlation between sequence length and execution time (AUCPreD Pearson's correlation coefficient of 0.71, metapredict Pearson's correlation coefficient of 0.88), yet the magnitude of the execution time for metapredict makes it look effectively flat. (B) Comparison of accuracy between AUCPreD and metapredict for the same sequences. The two methods are effectively comparable (see also Fig. 3A). (C) Zoomed-in comparison of execution time vs. protein size for metapredict. Note that the intercept here is 670 ms, which reflects the time needed to read-in and load the trained network, while the actual per-sequence execution time is under 100 ms for even a 1000 residue sequence (see also Supplemental Fig. 8 where execution times are calculated after the initial network file has been read in and parsed by metapredict).

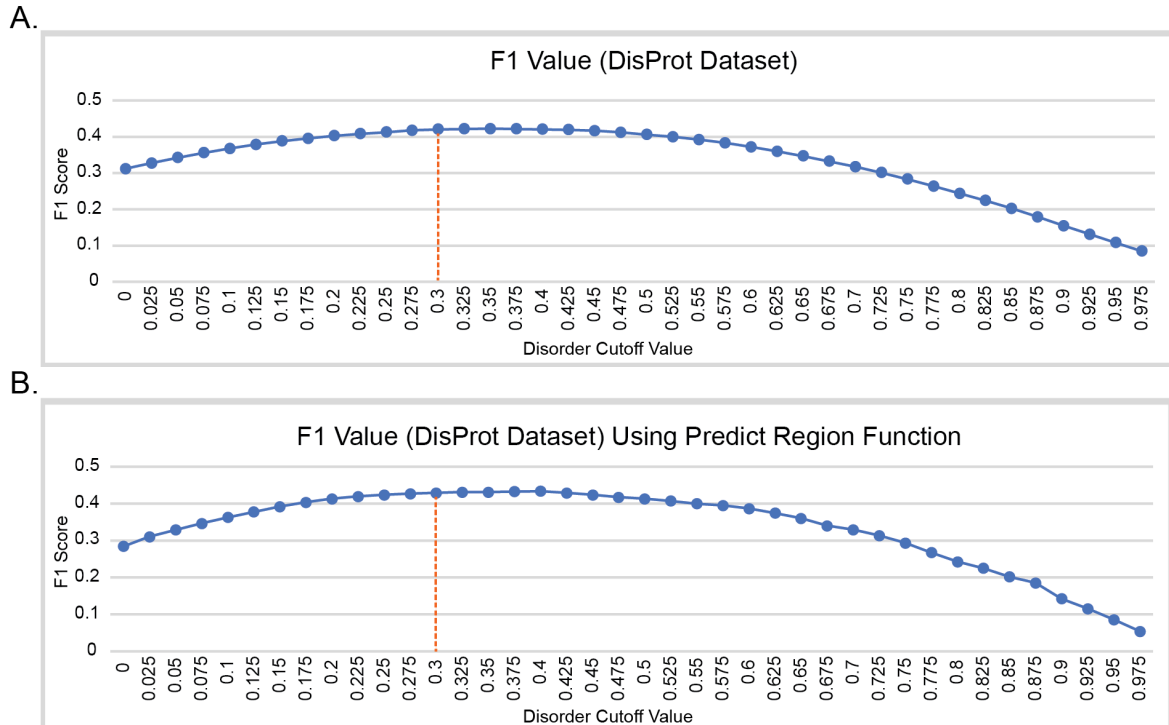


Supplemental Figure 3. Disorder and predicted structure confidence are highly correlated but independent measures of (lack of) protein structure.

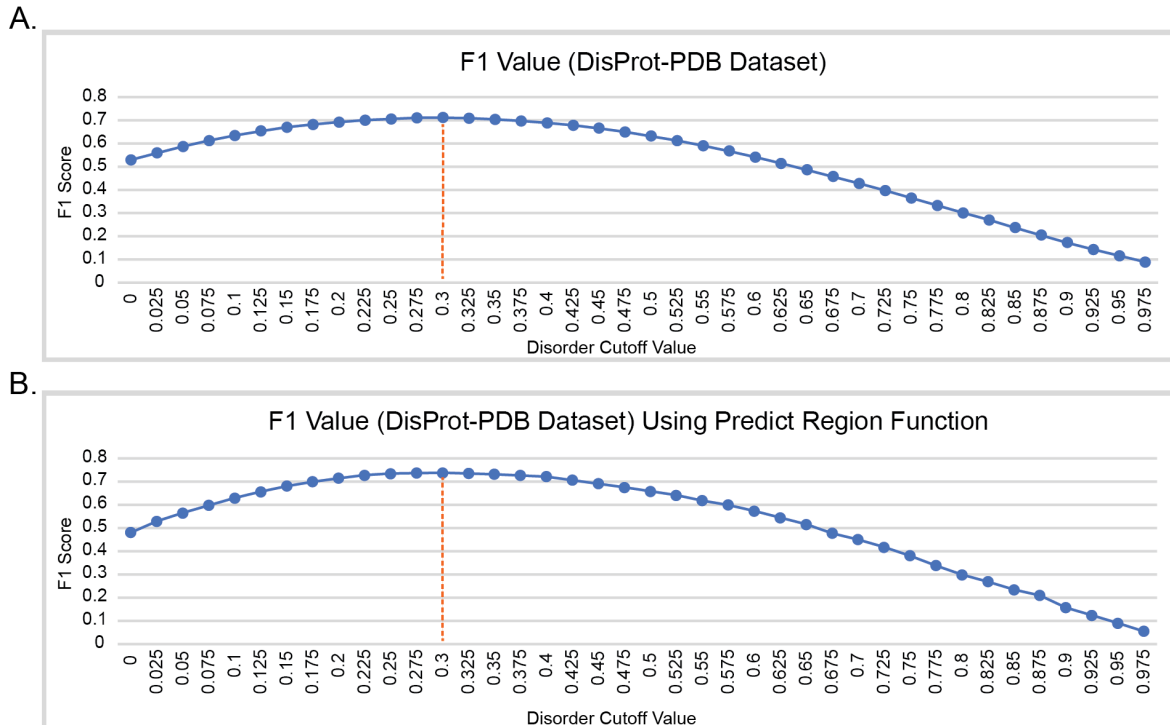
We analyzed every sequence in the human proteome, computed per-residue disorder predictions and predicted structure confidence scores (predicted pLDDT), and correlated the scores using the Pearson correlation coefficient. The histogram above represents the overall distribution of those correlation coefficients calculated for 20,394 protein sequences.



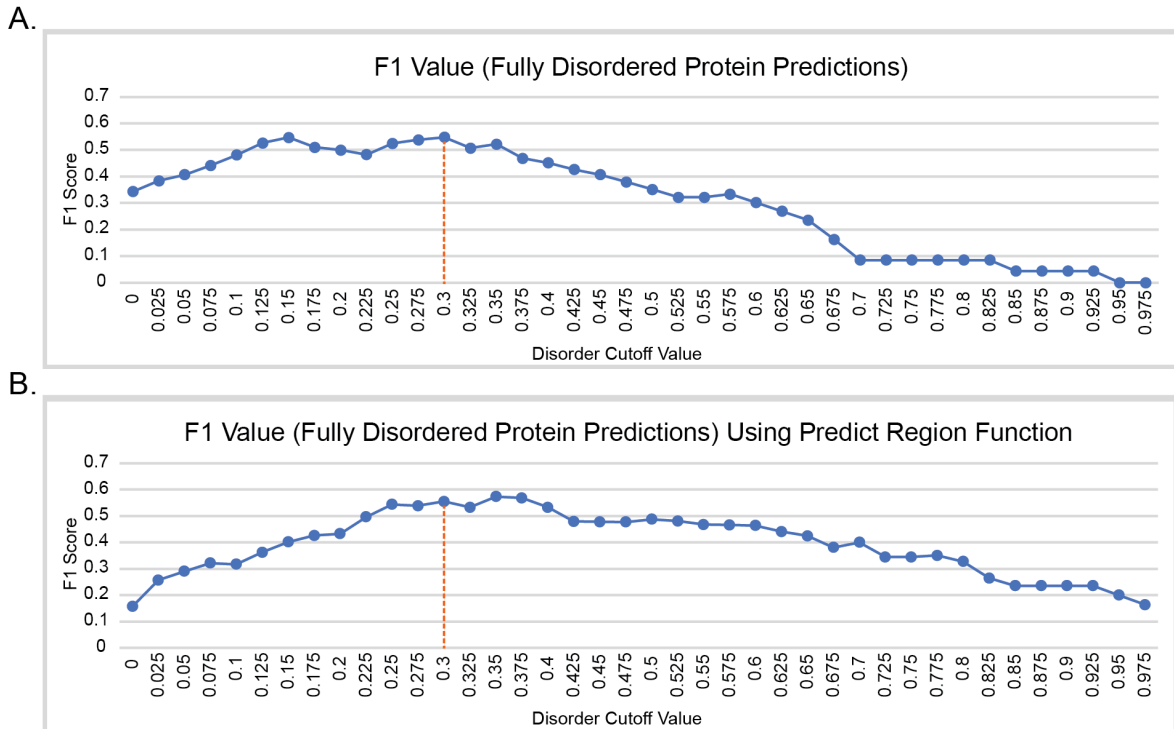
Supplemental Figure 4. Assessing the impact of disorder cutoff values on binary order and disorder classification of CheZOD data. (A) Absolute value of Pearson's Correlation Coefficient for binary predictions of disorder by metapredict compared to binary classifications of disorder from the CheZOD dataset. **(B)** Absolute value of Pearson's Correlation Coefficient for binary predictions of disorder by metapredict where the binary predictions were obtained using the `predict_disorder_domains()` function. These predictions are compared to binary classifications of disorder from CheZOD dataset. For both (A) and (B), the orange line represents the cutoff value used for binary classifications of order and disorder for metapredict.



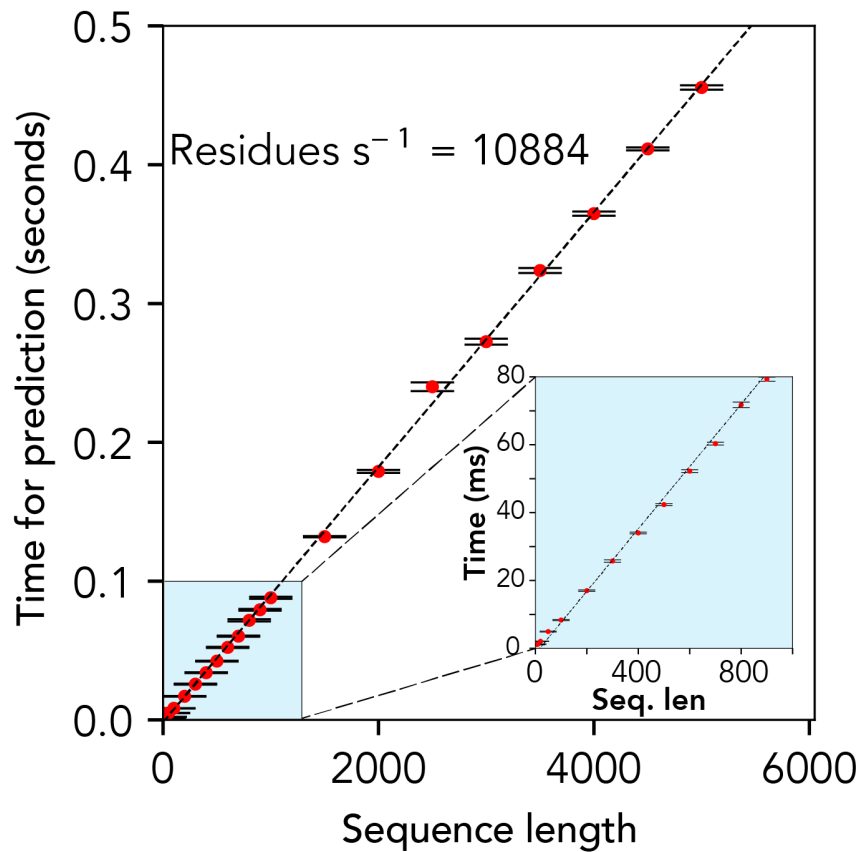
Supplemental Figure 5. Assessing the impact of disorder cutoff values on binary order and disorder classification of the Disprot Dataset (CAID). (A) F1-scores for binary predictions of disorder by metapredict compared to binary classifications of disorder from the Disprot dataset. (B) F1-scores for binary predictions of disorder by metapredict where the binary predictions were obtained using the `predict_disorder_domains()` function. These predictions were compared to binary classifications of disorder from the Disprot dataset. For both (A) and (B), the orange line represents the cutoff value used for binary classifications of order and disorder for metapredict.



Supplemental Figure 6. Assessing the impact of disorder cutoff values on binary order and disorder classification of the Disprot Dataset-PDB (CAID). (A) F1-scores for binary predictions of disorder by metapredict compared to binary classifications of disorder from the Disprot-PDB dataset. **(B)** F1-scores for binary predictions of disorder by metapredict where the binary predictions were obtained using the predict_disorder_domains() function. These predictions were compared to binary classifications of disorder from the Disprot-PDB dataset. For both (A) and (B), the orange line represents the cutoff value used for binary classifications of order and disorder for metapredict.



Supplemental Figure 7. Assessing the impact of disorder cutoff values on metapredict identifying fully disordered proteins from the Disprot Dataset (CAID). (A) F1-scores for predicted fully disordered proteins by metapredict compared to the known number of fully disordered proteins in the Disprot dataset. (B) F1-scores for predicted fully disordered proteins by metapredict where the binary predictions used to classify a protein as fully disordered were obtained using the predict_disorder_domains() function. These predictions were compared to the known number of fully disordered proteins in the Disprot dataset. For both (A) and (B), fully disordered proteins were counted if the predictor classified at least 95% of residues within a protein as disordered. For both (A) and (B), the orange line represents the cutoff value used for binary classifications of order and disorder for metapredict.



Supplemental Figure 8. Metapredict performance as a function of sequence length in number of residues. Assessment of length-dependence of metapredict performance reveals a linear scaling of prediction time with sequence length. Sequences here are randomly generated fixed-length sequences. Error bars are standard error of the mean calculated over thirty independent runs for random sequences of the specified length. Code for this analysis is provided in the Supporting Data GitHub repository.

A. metapredict in Python

```
import metapredict as meta

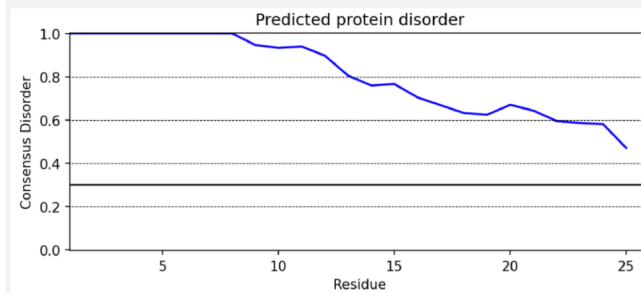
# define your sequence
my_seq = "MEEPQSDPSVEPPLSQETFSDLWKL"

# compute scores
disorder_scores = meta.predict_disorder(my_seq)

# we're done!
print(disorder_scores)

[1, 1, 1, 1, 1, 1, 1, 1, 0.946, 0.933, 0.939,
0.897, 0.805, 0.759, 0.766, 0.704, 0.668, 0.632,
0.624, 0.67, 0.643, 0.595, 0.586, 0.581, 0.471]

# create a plot of our disorder profile
meta.graph_disorder(my_seq)
```



B. metapredict in the terminal

```
$ ls
test_data.fasta
$ metapredict-predict-disorder test_data.fasta
$ ls
disorder_scores.csv test_data.fasta
```

C. metapredict online

metapredict online (v0.1)

metapredict is a deep-learning based consensus predictor of intrinsic disorder.

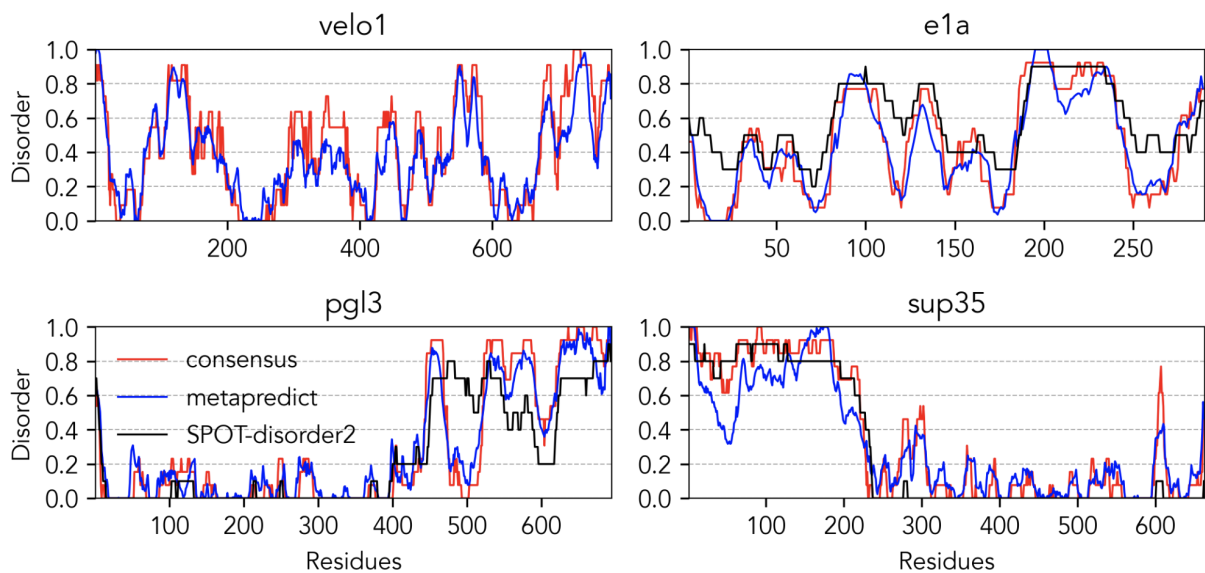
metapredict.net offers a simple online portal to some of metapredict's features but for a single sequence.

To use paste a single sequence into the input box below and select one of **plot disorder**, **get values**, or **get IDRs**.

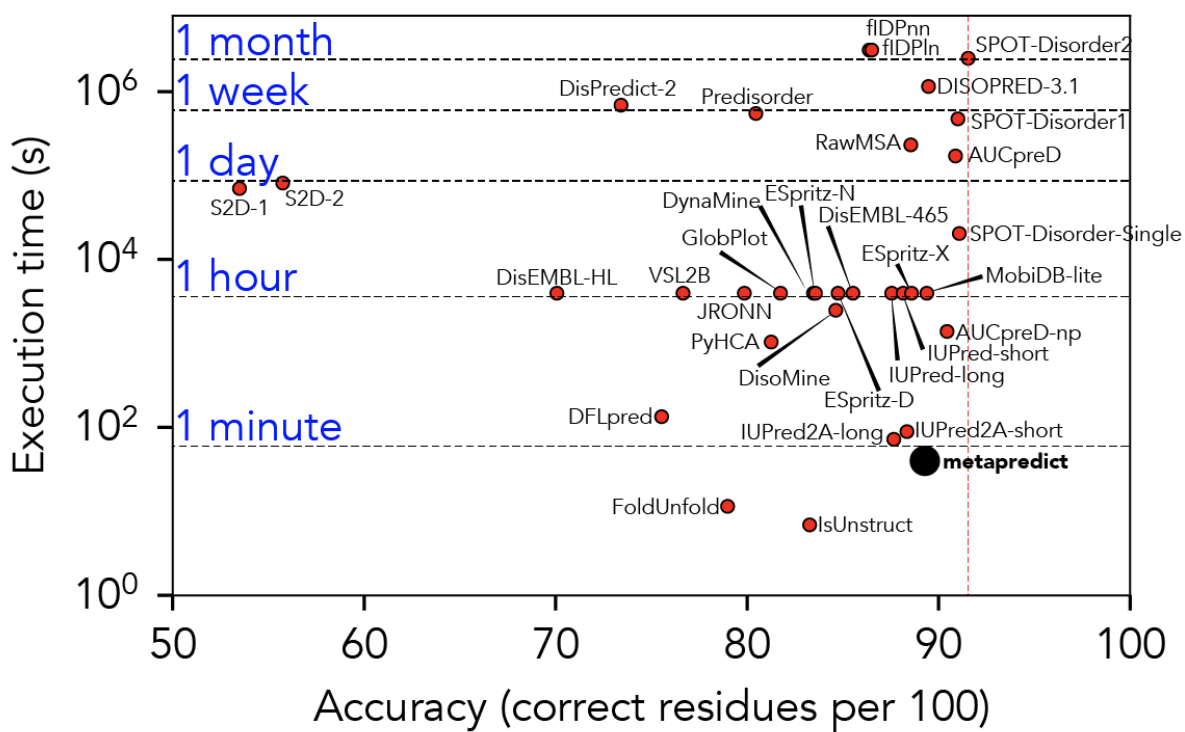
Generate a high-resolution, interactive and download-able plot of the per-residue disorder:

PLOT DISORDER

Supplemental Figure 9. Metapredict offers three distinct modes of use. (A) Metapredict can be used as a Python library, with simple and intuitive integration into existing Python code or for exploration in a Jupyter notebook. (B) Metapredict can be used as a command-line tool to interact directly with FASTA files. The file generated by the command metapredict-predict-disorder (“disorder_scores.csv”) is a simple comma separated value (CSV) file with per-residue disorder values provided for each sequence in the FASTA file. (C) Finally, metapredict is offered as a simple web server (<https://metapredict.net>), which can generate high-quality downloadable figures or allow per-residue disorder scores to be obtained as a CSV data file.



Supplemental Figure 10. Metapredict accurately recapitulates precomputed consensus disorder scores. Precomputed consensus disorder scores from the MobiDB database (red) and predicted disorder scores obtained SPOT-disorder2 (black) are compared to predicted consensus disorder scores calculated by metapredict for Velo1 from *Xenopus laevis* (UniProt Q7T226), PGL-3 from *Caenorhabditis elegans* (UniProt G5EBV6), Early E1A protein from Human adenovirus C serotype 5 (UniProt P03255), and Sup35 from *Schizosaccharomyces pombe* (UniProt O74718). None of these proteins were part of the training, test, or validation set for metapredicts. Note that Velo1 exceeds the length that SPOT-disorder2 can be used on.



Supplemental Figure 11. Reproduction of **Fig. 3B** with various predictors explicitly labelled.

Supplemental References

Acids Research, N., & 2021. (2021). UniProt: The universal protein knowledgebase in 2021.

Nucleic Acids Research, 49(D1), D480–D489.

Griffith, D., & Holehouse, A. S. (2021). PARROT: a flexible recurrent neural network framework for analysis of large protein datasets. In *bioRxiv* (p. 2021.05.21.445045).

<https://doi.org/10.1101/2021.05.21.445045>

Necci, M., Piovesan, D., CAID Predictors, DisProt Curators, & Tosatto, S. C. E. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*.

<https://doi.org/10.1038/s41592-021-01117-3>

Nielsen, J. T., & Mulder, F. A. A. (2019). Quality and bias of protein disorder predictors. *Scientific Reports*, 9(1), 1–11.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *arXiv [cs.LG]*. arXiv.

<http://arxiv.org/abs/1912.01703>

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. (2021). MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research*, 49(D1), D361–D367.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515.