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

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

Abstract

Intrinsically disordered proteins and protein regions make up a substantial fraction of many proteomes where they play a wide variety of essential roles. A critical first step in understanding the role of disordered protein regions in biological function is to identify those disordered regions correctly. Computational methods for disorder prediction have emerged as a core set of tools to guide experiments, interpret results, and develop hypotheses. Given the multiple different predictors available, consensus scores have emerged as a popular approach to mitigate biases or limitations of any single method. Consensus scores integrate the outcome of multiple independent disorder predictors and provide a per-residue value that reflects the number of tools that predict a residue to be disordered. Although consensus scores help mitigate the inherent problems of using any single disorder predictor, they are computationally expensive to generate. They also necessitate the installation of multiple different software tools, which can be prohibitively difficult. To address this challenge, we developed a deep-learning-based predictor of consensus disorder scores. Our predictor, metapredict, utilizes a bidirectional recurrent neural network trained on the consensus disorder scores from 12 proteomes. By benchmarking metapredict using two orthogonal approaches, we found that metapredict is among the most accurate disorder predictors currently available. Metapredict is also remarkably fast, enabling proteome-scale disorder prediction in minutes. Metapredict is fully open source and is distributed as a Python package, a collection of command-line tools, and a web server. We believe metapredict offers a convenient, accessible, accurate, and high-performance predictor for single-proteins and proteomes alike.

Introduction

While it is often convenient to consider proteins as nanoscopic molecular machines, such a description betrays many of their functionally critical features (1–3). As an extreme example, intrinsically disordered proteins and protein regions (collectively referred to as IDRs) do not adopt a fixed three-dimensional conformation (4–8). Rather, IDRs exist in an ensemble of different conformations that are in exchange with one another (9–11). Despite the absence of a well-defined structured state, IDRs are integral to many important biological processes (12, 13). As a result, there is a growing appreciation for the importance of disordered regions across the three kingdoms of life (6, 12, 14, 15).

A key first step in exploring the role of disorder in biological function is the identification of disordered regions. While IDRs can be formally identified by various biophysical methods (including nuclear magnetic resonance spectroscopy, circular dichroism, or single-molecule spectroscopy) these techniques can be challenging and are generally low throughput (16–18). As implied by the name, the “intrinsically” disordered nature of IDRs reflects the fact that these protein regions are unable to fold into a well-defined tertiary structure in isolation. This is in contrast to folded regions, which under appropriate solution conditions adopt macroscopically similar three-dimensional structures (19–21). The complexities of metastability in protein folding notwithstanding, this definition implies that this intrinsic ability to fold (or not fold) is encoded in the primary amino acid sequence (22–24). As such, it should be possible to delineate between folded and disordered regions based solely on amino acid sequence.

The prediction of protein disorder from amino acid sequence has received considerable attention for over twenty years, driven by pioneering early work by Dunker *et al*. (6–8, 25, 26). Since those original bioinformatics tools, a wide range of disorder predictors have emerged (27–30). Accurate disorder predictors offer an approach to guide experimental design, interpret data, and build testable hypotheses. As such, the application of disorder predictors to assess predicted protein structure has become a relatively standard type of analysis, although the specific predictor used varies depending on availability, simplicity, and scope of the question.

There are currently many disorder predictors that apply different approaches to predict protein disorder. These range from statistical approaches based on structural data from the protein data bank, to biophysical methods that consider local ‘foldability’, to machine learning-based algorithms trained on experimentally determined disordered sequences (31–38). However, using any individual predictor can be problematic; each predictor has specific biases and weaknesses in its capacity to accurately predict protein disorder, which can introduce systematic biases into large-scale disorder assessment (39). As such, an alternative strategy in which many different predictors are combined to offer a consensus disorder score has emerged as a popular alternative to relying on any specific predictor (40–44). Consensus scores report the fraction of independent disorder predictors that would predict a given residue as disordered - for example, a score of 0.5 reports that 50% of predictors predict that residue to be disordered.

While using consensus scores mitigates the limitations of any single predictor, calculating consensus scores is computationally expensive. On a practical level, the installation and operation of many disorder predictors has specific and often complicated requirements, each of which may also be operating system specific. This raises a significant barrier to computational and non-computational scientists alike. To alleviate this challenge, consensus disorder scores can be precomputed and held in online-accessible databases (42, 45–47). While precomputed scores are an invaluable resource to the scientific community, they are limited to a set of specific protein sequences. This raises a challenge - while consensus scores are a convenient means by which to explore protein disorder, their application is limited to a small subset of possible sequences. Furthermore, obtaining, managing, and analyzing large datasets of precomputed consensus predictions can be a daunting task, especially if only a subset of sequences are of interest.

To address these challenges we have developed a fast, accurate, and simple-to-use deep learning-based disorder predictor trained on pre-computed consensus scores from a range of organisms. Our resulting predictor, metapredict, is platform agnostic, simple to install, and usable as both a Python module and a stand-alone command-line tool. In addition, metapredict is implemented in a stand-alone web server, appropriate for individual sequences. Metapredict accurately reproduces consensus disorder scores and is sufficiently fast that for most bioinformatics pipelines, precomputation of disorder is no longer necessary, and disorder can be computed in real-time as analysis is performed. Taken together, metapredict offers a lightweight, accurate, and high-performance disorder predictor that can be readily integrated into existing workflows.

Materials and Methods

Training metapredict using PARROT

To create metapredict, we trained a bidirectional recurrent neural network (BRNN) on the disorder consensus scores from the MobiDB database for each residue for all of the proteins in 12 proteomes (**Figure 1**) (48). The eight disorder predictors used to generate the consensus scores in the MobiDB database were IUPred short (34), IUPred long (34), ESpiritz (DisProt, NMR, and X-ray) (31), DisEMBL 465 (28), DisEMBL hot loops (28), and GlobPlot (49). For training, we used the general-purpose deep learning toolkit PARROT to train a Long Short-Term Memory (LSTM) BRNN (50). PARROT offers a general flexible framework for the development and training of deep-learning-based predictors for sequence data.

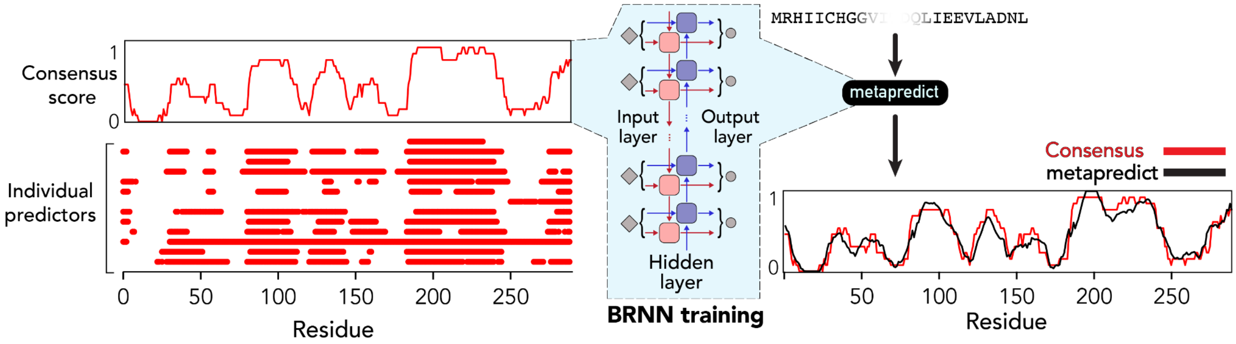


Figure 1. Overview of metapredict. Consensus scores are taken from 420,660 proteins distributed across 12 proteomes. Metapredict was developed by training a BRNN on this data, leading to a set of network weights that allow the prediction of any possible consensus sequence score.

Specifically, we used PARROT with slight modifications to the default settings (50). 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 (42, 51, 52).

To determine the optimal threshold used to delineate disordered and ordered regions, we systematically varied the cutoff score used to identify IDRs (**Supplemental Figures 1-4**). This analysis revealed that a relatively broad range of cutoffs (between 0.2 and 0.4) gave an approximately equivalent performance, such that a cutoff of 0.3 offered a good balance between true positives and false negatives. As such, IDRs identified by metapredict with the default setting can be treated as relatively high-confidence, at the expense of missing some cryptic disordered regions.

Evaluating metapredict

Evaluations and datasets for the CheZOD score analysis (116 sequences) can be found in (53). Evaluations and datasets for the Critical Assessment of protein Intrinsic Disorder prediction (CAID) analysis (652 sequences) can be found in (27). For convenience, all sequences and scores used are also provided at https://github.com/holehouse-lab/supportingdata/. 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 **Supplemental File 1**.

Statistical analysis

Statistical analysis and predictor evaluation was carried out following the protocols described previously and reproduced here for completeness (27, 53). 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,

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

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,

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

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

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

Where TP, FP, and FN are defined as described above.

Implementation

metapredict is written in Python 3.7+ and uses PyTorch, with the initial network trained using PARROT (50, 54).

Usage and features

metapredict is offered in three distinct formats. As a downloadable package, it can be used either as via a set of command-line tools or as a Python module. Command-line predictions include functionality to directly predict disorder from a UniProt accession, save disorder scores as a text file, and predict disorder for multiple sequences within a FASTA file. The Python module includes the ability to predict per-residue consensus disorder scores or delineate continuous IDRs. Complete documentation is available at http://metapredict.readthedocs.io/. In addition, we offer a web server appropriate for individual protein sequences, which is available at http://metapredict.net.

Performance

On all hardware tested, metapredict obtained prediction rates of ~8,000 to 12,000 residues per second. At this performance, a single 300-residue protein takes between 20-30 ms, and the complete prediction of all IDRs across the reviewed human proteome (20,396 sequences) takes approximately 21 minutes with an average performance of ~9,000 residues per second. Importantly and unlike some other predictors, the computational cost scales linearly with sequence length (**Supplemental figure 5**) (55). For reference, disorder prediction for the complete CAID dataset takes ~30 seconds.

Results

Approaches to evaluating metapredict

Given the large number of protein disorder predictors available, multiple groups have investigated different approaches by which to measure their accuracy (27, 53, 56, 57). Here, we used metrics from two recent studies, allowing us to compare directly with many previously evaluated predictors.

The first approach made use of data obtained from nuclear magnetic resonance (NMR) spectroscopy, specifically a metric referred to as the Chemical shift Z-score for assessing Order/Disorder (CheZOD) score (53). The CheZOD score describes how far the measured chemical shift of a given residue deviates from the expected random coil chemical shifts. As a result, CheZOD offers a continuous score at the resolution of individual residues (53).

The second approach assessed how well various disorder predictors were able to predict the disorder of proteins from the DisProt database (27, 58). In this approach, regions deposited in the DisProt database were considered true positives, providing a simple test set against which disorder predictors could be evaluated.

Evaluating metapredict using CheZOD scores

To evaluate metapredict using the CheZOD Z-scores, we carried out the same analyses that had previously been done to examine other predictors using the same set of proteins as in (53). In these analyses, the authors examined how the continuous probability values of the predictors correlated with the continuous values of the Z-scores using the Pearson correlation coefficient (Eq. 1). Because the CheZOD Z-scores increase with order and decrease with disorder, a score of -1 for a disorder predictor would mean that it has a perfect correlation 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. Using this analysis, we found that metapredict had ranked 8th among the 23 predictors previously examined (**Figure 2A**). In addition to examining the accuracy of metapredict using the absolute value of the Pearson Correlation coefficient, we also calculated the area under a receiver operating characteristic curve, which 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. In examining metapredict using this approach, we found that metapredict was ranked 11th out of the 23 predictors evaluated (**Figure 2B**).

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 (53). 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 File 1**). In examining where metapredict did poorly in classifying residues as either ordered or disordered, we found that metapredict suffered from a relatively high false-negative rate suggesting that metapredict was classifying residues as ordered when they were in fact disordered.

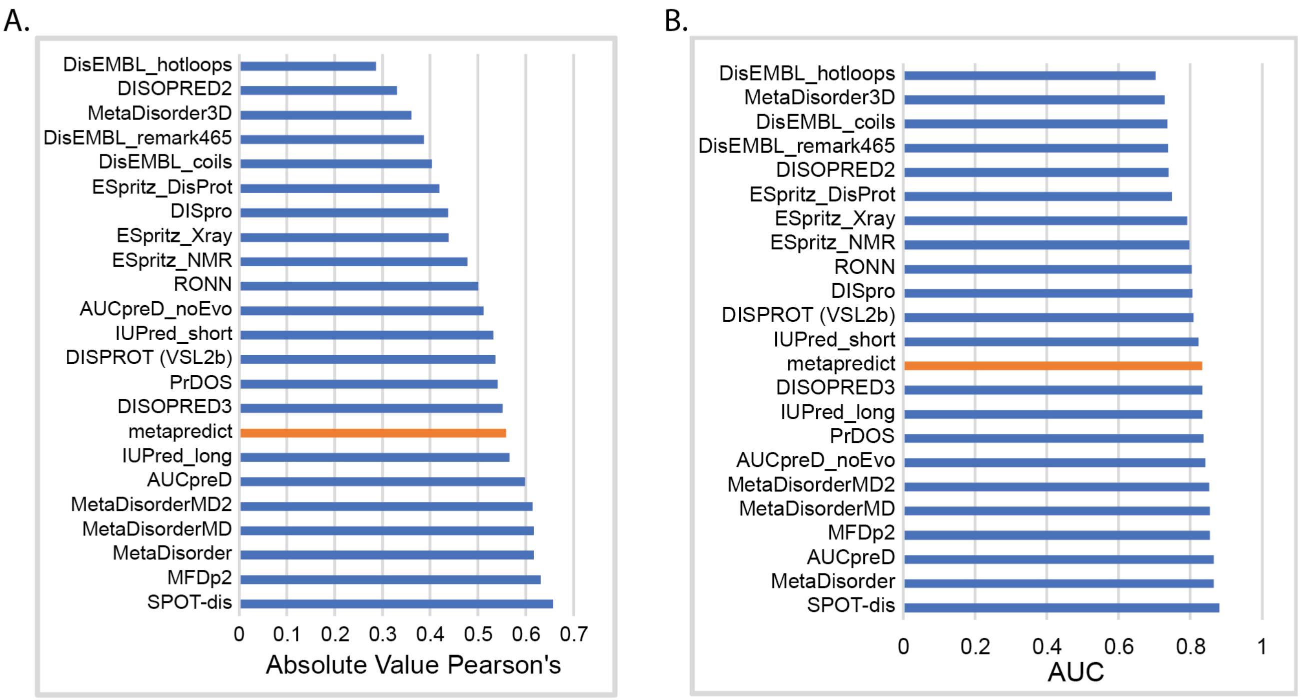


Figure 2. Evaluation of metapredict using CheZOD scores. (A) The absolute value of the Pearson's correlation coefficient calculated by comparing the correlation between each predictor's score per residue and the CheZOD score. **(B)** The area under the receiver operating characteristic curve (AUC) (generated 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 (53).

Evaluating metapredict using the Critical Assessment of protein Intrinsic Disorder prediction

We next turned to evaluate metapredict using the protocol developed for the Critical Assessment of protein Intrinsic Disorder prediction (CAID). CAID is a biennial event in which a large set of protein disorder predictors are assessed using a standardized dataset and standardized metrics (27). As such, evaluation using CAID's standards offers a convenient route to benchmark metapredict against the state of the art.

CAID uses a curated dataset of 646 proteins from DisProt, a database of experimentally validated disordered regions (58). CAID also standardizes hardware upon which predictions are run, allowing performance and runtime to be considered with respect to predictor performance. To avoid inadvertently misrepresenting our results with respect to performance through the use of non-standard hardware, we focused on a subset of the CAID protocol focused on the accuracy of predictors, although we note metapredict's computational efficiency performance is a major strength.

In keeping with the assessments developed by CAID, we evaluated metapredict in its capacity to predict disorder across two distinct databases (DisProt, DisProt-PDB) as well as its ability to identify fully disordered proteins (27). While DisProt contains only true positive disordered regions, DisProt-PDB contains true positive and true negative regions, making it more appropriate for robust validation of discriminatory predictors (27). To maintain consistency with CAID, we used the F1-score (defined as the maximum harmonic mean between precision and recall across all threshold values, Eq. 3) to compare metapredict against other predictors (27). The F1-score of metapredict in the analysis of the DisProt dataset ranked 12th highest out of the 38 predictors originally assessed (**Figure 3A**).

DisProt contains protein subregions that have been experimentally validated as disordered. However, as noted in the original study, it is possible, if not likely, that there are other subregions from those same proteins which, while not yet annotated as such, are in fact disordered (27). The DisProt-Protein Database (PDB) dataset addresses this limitation and includes only protein regions that are unambiguously annotated as either disordered or ordered, based on extant experimental data (27). In examining the performance of metapredict in predicting disorder on the DisProt-PDB dataset, we found that metapredict ranked 11th among all of the disorder predictors assessed (**Figure 3B**).

The last analysis that we carried out from the CAID experiment was the capacity of metapredict to identify fully disordered proteins. In this context, the CAID experiment considers something to be a fully disordered protein if the disorder predictor predicts 95% or more residues to be disordered (27). Metapredict ranked 3rd out of the disorder predictors examined in its capacity to identify fully disordered proteins (**Figure 3C**).

In examining the results from the three CAID assessments more closely, we found that in comparison to the other disorder predictors, metapredict was most negatively impacted by its high frequency of false-negative predictions (where metapredict predicted something to be ordered when it was in fact disordered). This is consistent with our comparison against the CheZOD scores. As such, metapredict appears to possess a slight bias towards underestimating disorder, such that IDRs identified by metapredict can be considered reasonably high confidence.

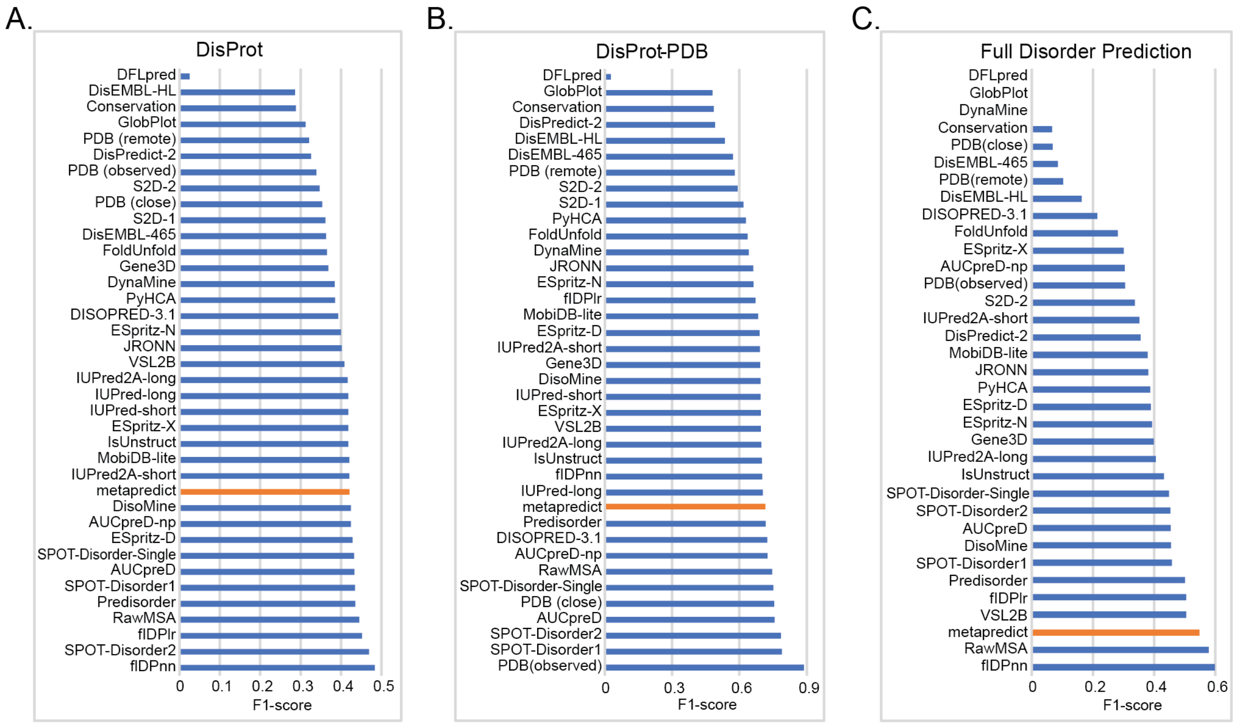


Figure 3. Evaluation of metapredict using CAID experiments. (A) F1-score for various predictors in examining their accuracy in predicting protein disorder from the DisProt dataset. **(B)** F1-scores for various predictors in examining their accuracy in predicting protein disorder from the DisProt-PDB dataset. **(C)** F1-scores for various predictors in predicting fully disordered proteins in the DisProt dataset. Values for all predictors in (A), (B), and (C) with the exception of those for metapredict (orange bar) were obtained from (27).

Discussion

IDRs play many important roles in various biological processes (12, 13). An essential first step in the investigation of IDR function reflects the ability to identify IDRs within a protein sequence. Consensus disorder scores represent an attractive means by which to obtain high confidence disorder predictions that do not suffer from inaccuracies due to the limitations of any single disorder predictor. However, calculating disorder probabilities from many different predictors to generate a consensus score is cumbersome, technically challenging, and computationally expensive. To address this, we developed metapredict, a simple to use protein disorder predictor that accurately reproduces consensus disorder scores. While other consensus meta-predictors do exist, web-based access to these can be on the order of minutes-to-hours per sequence and where available local access has operating-system dependencies, making them poorly suited to cross-platform proteome-scale analysis (41, 59, 60). As such, we believe metapredict fills a niche that is currently unoccupied.

To illustrate the ability of metapredict to predict consensus scores, **Figure 4** 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.

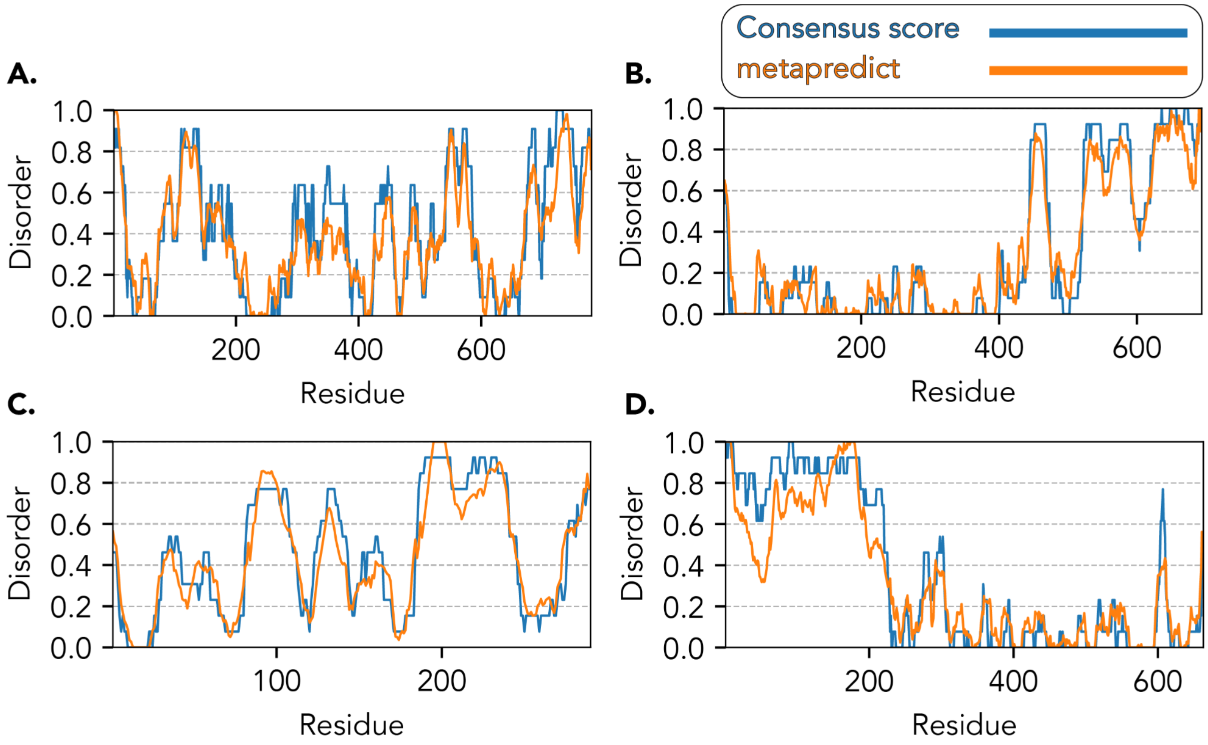


Figure 4. Metapredict accurately recapitulates precomputed consensus disorder scores. Precomputed consensus disorder scores from the MobiDB database compared to predicted consensus

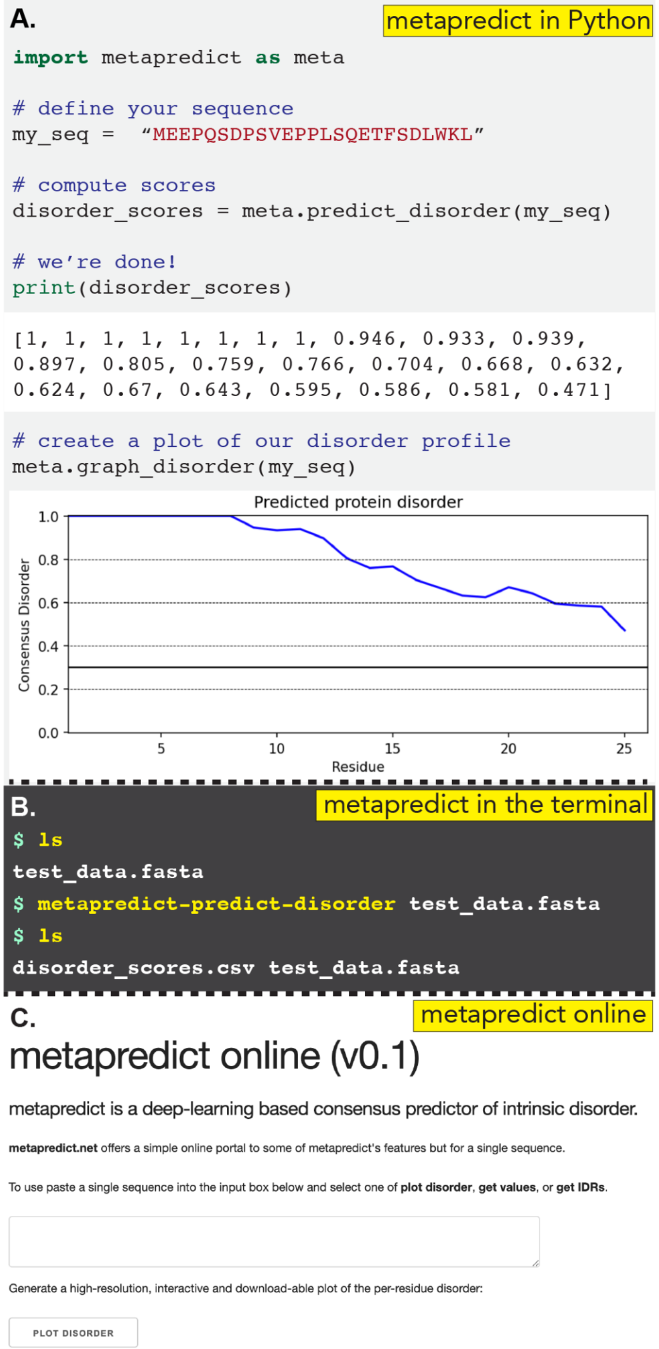
disorder scores calculated by metapredict for **(A)** Velo1 from *Xenopus laevis* (UniProt Q7T226), **(B)** PGL-3 from *Caenorhabditis elegans* (UniProt G5EBV6), **(C)** Early E1A protein from Human adenovirus C serotype 5 (UniProt P03255), and **(D)** Sup35 from *Schizosaccharomyces pombe* (UniProt O74718). None of these proteins were part of the training, test, or validation set for metapredicts.

Features of metapredict

While metapredict did not score as the most accurate available protein disorder predictor based on the analyses presented here, it was still consistently among the most accurate predictors. Furthermore, the difference in F-score between metapredict and the best-scoring predictor across the CAID dataset was on average only 0.095. We would propose that the convenience, computational efficiency, and ease-of-use that metapredict affords outweigh what amounts to a relatively small cost in overall accuracy. Similarly, metapredict's flexibility, ease of installation, and lack of dependencies beyond standard Python packages make it an appealing option for modern disorder prediction across heterogeneous hardware.

To further aid in the identification of *bona fide* contiguous disordered regions, metapredict contains a stand-alone function for extracting contiguous IDRs based on a threshold value applied to a smoothed disorder score and several additional parameters (**Supplemental Figures 1-4**). For this approach, we again found a threshold between 0.3 and 0.4 was optimal, and this method generally outperformed our prior more simple analyses. However, because other predictors did not use this approach for the classification of ordered or disordered regions, we chose to not use this function for our primary analyses in examining the accuracy of metapredict. Nonetheless, this suggests that metapredict can achieve even marginally higher accuracy in identifying IDRs, and automates this procedure for the users, allowing boundaries between IDRs and folded domains to be automatically identified and greatly facilitating IDR-ome style analyses of datasets.

We designed metapredict to be as flexible and user-friendly as possible. For example, metapredict can be used as a Python library (**Figure 5A),** a stand-alone command-line tool (**Figure 5B**) or a web server (http://metapredict.net) (**Figure 5C**). Moreover, metapredict contains functionality to generate graphs or disorder scores from the command-line by directly inputting a single protein sequence or even by passing a UniProt accession number. Alternatively, the user can pass in protein sequences in a FASTA formatted file, and metapredict will return either raw scores or individual graphs for each protein. 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 the 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.



```
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)
```

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

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:

Figure 5. 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.

Code and data availability

The code for metapredict can be found at: https://github.com/idptools/metapredict. Documentation is available at https://metapredict.readthedocs.io/. Fully processed sequences used for assessment (including sequences and scores) are provided at https://github.com/holehouse-lab/supportingdata/. Metapredict can be installed directly from the Python Packaging Index using pip (i.e., pip install metapredict).

Author Contributions

R. J. E. designed research, developed code, performed analysis, made figures, and wrote the manuscript. D.G. developed code and edited the manuscript. A.S.H. designed research, developed code, made figures, and wrote the manuscript.

Acknowledgments

Funding for this project was provided by the Longer Life Foundation (an RGA/Washington University Collaboration) to A.S.H. and the William H. Danforth Plant Science Fellowship to R.J.E. We thank Ishan Taneja and Jeff Lotthammer for helpful comments on the manuscript. We thank Steven Boeynaems for the “motivation” to develop our web server. We also thank the entire Tosatto group, the ELIXIR Intrinsically Disordered Proteins Community, and HUPO-PSI Intrinsically Disordered Proteins Community (notably Silvio Tosatto, Zsuzsanna Dosztanyi, Damiano Piovesan, Wim Vranken and Norman Davey) for all the European-funded bioinformatics work that largely fuels the international IDP informatics space.

References

1. Sormanni, P., D. Piovesan, G.T. Heller, M. Bonomi, P. Kukic, C. Camilloni, M. Fuxreiter, Z. Dosztanyi, R.V. Pappu, M.M. Babu, S. Longhi, P. Tompa, A.K. Dunker, V.N. Uversky, S.C.E. Tosatto, and M. Vendruscolo. 2017. Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* 13:339–342.

2. Bottaro, S., and K. Lindorff-Larsen. 2018. Biophysical experiments and biomolecular simulations: A perfect match? *Science*. 361:355–360.

3. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. *Nature*. 450:964–972.

4. van der Lee, R., M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R.V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, and M.M. Babu. 2014. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* 114:6589–6631.

5. Wright, P.E., and H.J. Dyson. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293:321–331.

6. Dunker, A.K., Z. Obradovic, P. Romero, E.C. Garner, and C.J. Brown. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* 11:161–171.

7. Uversky, V.N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11:739–756.

8. Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem. Sci.*

9. Mittag, T., and J.D. Forman-Kay. 2007. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* 17:3–14.

10. Forman-Kay, J.D., and T. Mittag. 2013. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure*. 21:1492–1499.

11. Mao, A.H., N. Lyle, and R.V. Pappu. 2013. Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem. J*. 449:307–318.

12. Wright, P.E., and H.J. Dyson. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16:18–29.

13. Oldfield, C.J., and A.K. Dunker. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* 83:553–584.

14. Tompa, P., and M. Fuxreiter. 2008. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.*

15. Tompa, P., and A. Fersht. 2009. Structure and Function of Intrinsically Disordered Proteins. CRC Press.

16. Gibbs, E.B., E.C. Cook, and S.A. Showalter. 2017. Application of NMR to studies of intrinsically disordered proteins. *Arch. Biochem. Biophys.* 628:57–70.

17. Chemes, L.B., L.G. Alonso, M.G. Noval, and G. de Prat-Gay. 2012. Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. *Methods Mol. Biol.* 895:387–404.

18. Schuler, B., A. Soranno, H. Hofmann, and D. Nettels. 2016. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* 45:207–231.

19. Karplus, M., and D.L. Weaver. 1976. Protein-folding dynamics. *Nature*. 260:404–406.

20. Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science*. 181:223–230.

21. Dill, K.A., and H.S. Chan. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19.

22. Honeycutt, J.D., and D. Thirumalai. 1990. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* 87:3526–3529.

23. Thirumalai, D., and G. Reddy. 2011. Are native proteins metastable? *Nat. Chem.* 3:910–911.

24. Hu, X., L. Hong, M. Dean Smith, T. Neusius, X. Cheng, and J.C. Smith. 2016. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nat. Phys.* 12:171–174.

25. Romero, P., Z. Obradovic, C. Kissinger, J.E. Villafranca, and A.K. Dunker. 1997. Identifying disordered regions in proteins from amino acid sequence. In: Proceedings of International Conference on Neural Networks (ICNN'97). . pp. 90–95 vol.1.

26. Romero, Obradovic, and K. Dunker. 1997. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform. Ser. Workshop Genome Inform.* 8:110–124.

27. Necci, M., D. Piovesan, CAID Predictors, DisProt Curators, and S.C.E. Tosatto. 2021. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*.

28. Linding, R., L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, and R.B. Russell. 2003. Protein disorder prediction: implications for structural proteomics. *Structure*. 11:1453–1459.

29. Ferron, F., S. Longhi, B. Canard, and D. Karlin. 2006. A practical overview of protein disorder prediction methods. *Proteins*. 65:1–14.

30. Deng, X., J. Eickholt, and J. Cheng. 2012. A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.* 8:114–121.

31. Walsh, I., A.J.M. Martin, T. Di Domenico, and S.C.E. Tosatto. 2012. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*. 28:503–509.

32. Mészáros, B., G. Erdős, and Z. Dosztányi. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46:W329–W337.

33. Dosztányi, Z., V. Csizmók, P. Tompa, and I. Simon. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 347:827–839.

34. Dosztányi, Z., V. Csizmok, P. Tompa, and I. Simon. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 21:3433–3434.

35. Dass, R., F.A.A. Mulder, and J.T. Nielsen. 2020. ODiNPred: comprehensive prediction of protein order and disorder. *Sci. Rep.* 10:1–16.

36. Hanson, J., K.K. Paliwal, T. Litfin, and Y. Zhou. 2019. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics Proteomics Bioinformatics*. 17:645–656.

37. Ishida, T., and K. Kinoshita. 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35:W460–4.

38. Mizianty, M.J., Z. Peng, and L. Kurgan. 2013. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord Proteins*. 1:e24428.

39. Katuwawala, A., C.J. Oldfield, and L. Kurgan. 2020. Accuracy of protein-level disorder predictions. *Brief. Bioinform.* 21:1509–1522.

40. Necci, M., D. Piovesan, Z. Dosztányi, and S.C.E. Tosatto. 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*. 33:1402–1404.

41. Kozlowski, L.P., and J.M. Bujnicki. 2012. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*. 13:1–11.

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

43. Necci, M., D. Piovesan, D. Clementel, Z. Dosztányi, and S.C.E. Tosatto. 2020. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinformatics*.

44. Peng, Z., and L. Kurgan. 2012. On the complementarity of the consensus-based disorder prediction. *Pac. Symp. Biocomput.* 176–187.

45. Di Domenico, T., I. Walsh, and S.C.E. Tosatto. 2013. Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. *BMC Bioinformatics*. 14 Suppl 7:S3.

46. Oates, M.E., P. Romero, T. Ishida, M. Ghalwash, M.J. Mizianty, B. Xue, Z. Dosztányi, V.N. Uversky, Z. Obradovic, L. Kurgan, A.K. Dunker, and J. Gough. 2013. D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* 41:D508–16.

47. Potenza, E., T. Di Domenico, I. Walsh, and S.C.E. Tosatto. 2015. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* 43:D315–D320.

48. Piovesan, D., F. Tabaro, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A.M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W.F. Vranken, and S.C.E. Tosatto. 2018. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* 46:D471–D476.

49. Linding, R., R.B. Russell, V. Neduva, and T.J. Gibson. 2003. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31:3701–3708.

50. Griffith, D., and A.S. Holehouse. 2021. PARROT: a flexible recurrent neural network framework for analysis of large protein datasets. *bioRxiv*. 2021.05.21.445045.

51. Acids Research, N., and 2021. 2021. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489.

52. UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47:D506–D515.

53. Nielsen, J.T., and F.A.A. Mulder. 2019. Quality and bias of protein disorder predictors. *Sci. Rep.* 9:1–11.

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

55. Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio. 2016. Deep learning. MIT press Cambridge.

56. Monastyrskyy, B., K. Fidelis, J. Moult, A. Tramontano, and A. Kryshtafovych. 2011. Evaluation of disorder predictions in CASP9. *Proteins*. 79 Suppl 10:107–118.

57. Monastyrskyy, B., A. Kryshtafovych, J. Moult, A. Tramontano, and K. Fidelis. 2014. Assessment of protein disorder region predictions in CASP10. *Proteins*. 82 Suppl 2:127–137.

58. Hatos, A., B. Hajdu-Soltész, A.M. Monzon, N. Palopoli, L. Álvarez, B. Aykac-Fas, C. Bassot, G.I. Benítez, M. Bevilacqua, A. Chasapi, L. Chemes, N.E. Davey, R. Davidović, A.K. Dunker, A. Elofsson, J. Gobeill, N.S.G. Foutel, G. Sudha, M. Guharoy, T. Horvath, V. Iglesias, A.V. Kajava, O.P. Kovacs, J. Lamb, M. Lambrughi, T. Lazar, J.Y. Leclercq, E. Leonardi, S. Macedo-Ribeiro, M. Macossay-Castillo, E. Maiani, J.A. Manso, C. Marino-Buslje, E. Martínez-Pérez, B. Mészáros, I. Mičetić, G. Minervini, N. Murvai, M. Necci, C.A. Ouzounis, M. Pajkos, L. Paladin, R. Pancsa, E. Papaleo, G. Parisi, E. Pasche, P.J. Barbosa Pereira, V.J. Promponas, J. Pujols, F. Quaglia, P. Ruch, M. Salvatore, E. Schad, B. Szabo, T. Szaniszló, S. Tamana, A. Tantos, N. Veljkovic, S. Ventura, W. Vranken, Z. Dosztányi, P. Tompa, S.C.E. Tosatto, and D. Piovesan. 2020. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 48:D269–D276.

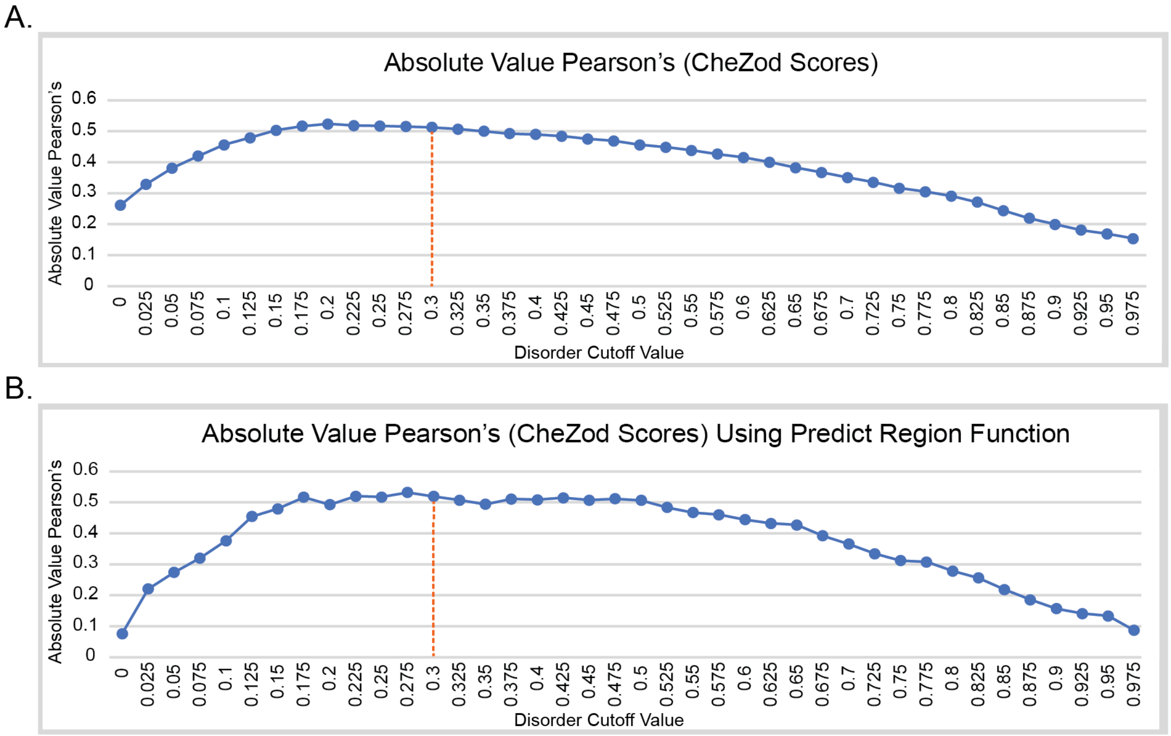
59. Schlessinger, A., M. Punta, G. Yachdav, L. Kajan, and B. Rost. 2009. Improved disorder prediction by combination of orthogonal approaches. *PLoS One*. 4:e4433.

60. Xue, B., R.L. Dunbrack, R.W. Williams, A.K. Dunker, and V.N. Uversky. 2010. PONDR-FIT:

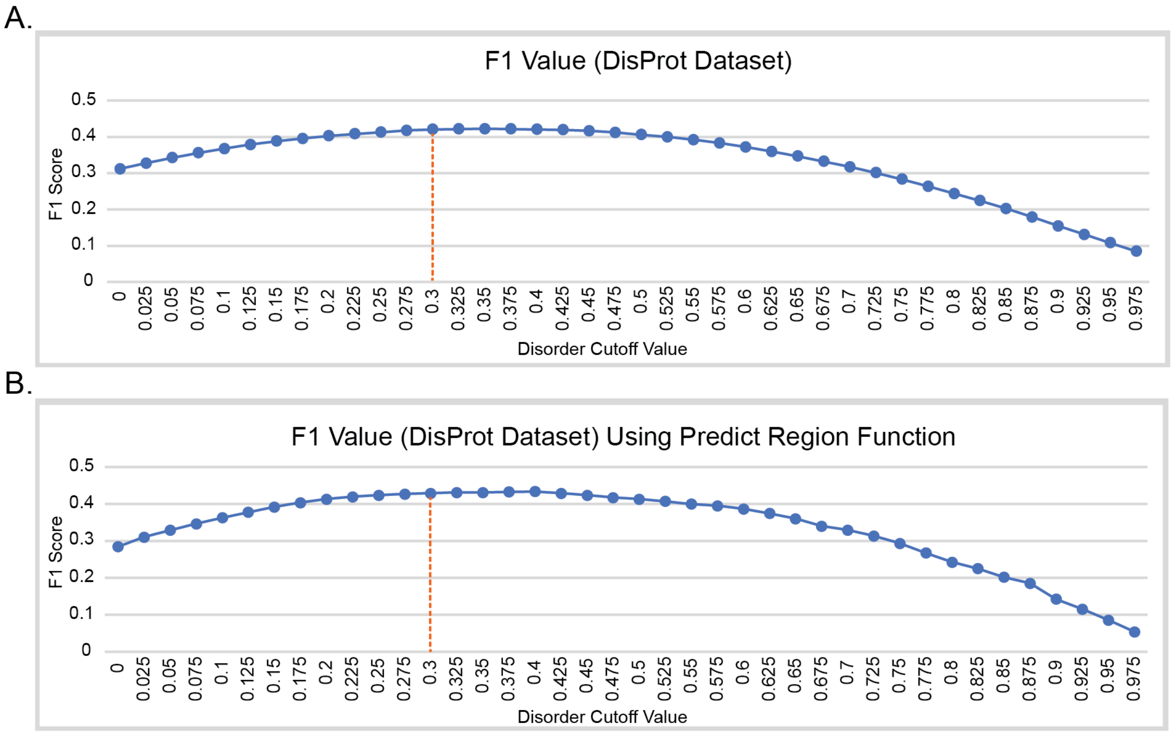
a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*. 1804:996–1010.

Supplemental Information

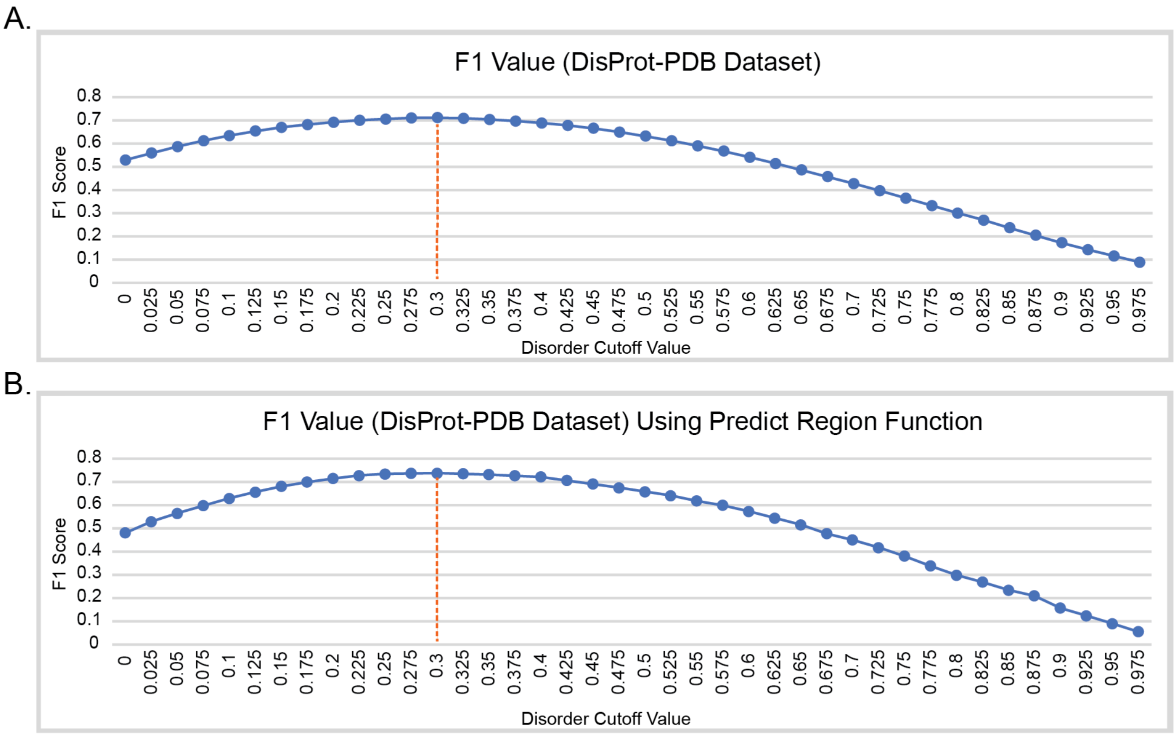
Supplemental File 1 - Summary of all analyses including additional statistical analyses not shown in the primary manuscript as well as raw values for each analysis for all predictors examined. All values with the exception of those for metapredict were obtained from (27, 53).



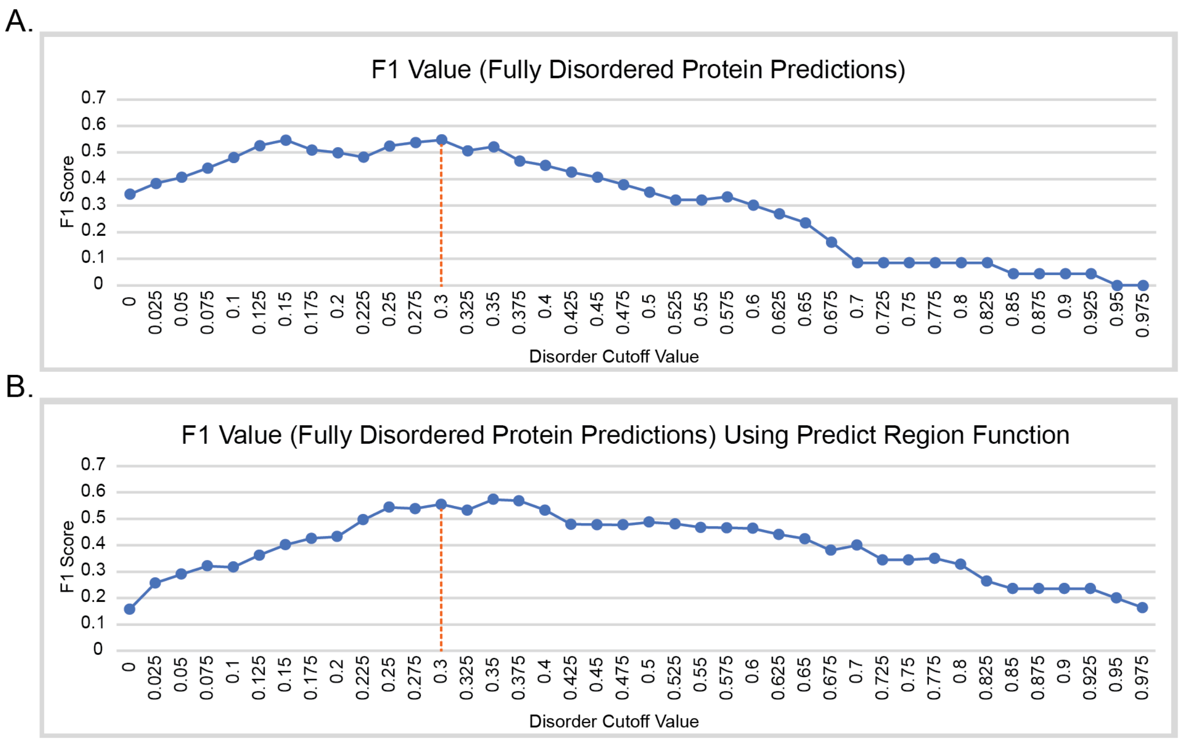
Supplemental Figure 1. Assessing the impact of disorder cutoff values on binary order and disorder classification of CheZOD data. (A) Absolute value of Pearson's Correlation Coefficient for binary predictions of disorder by metapredict compared to binary classifications of disorder from the CheZOD dataset. **(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.



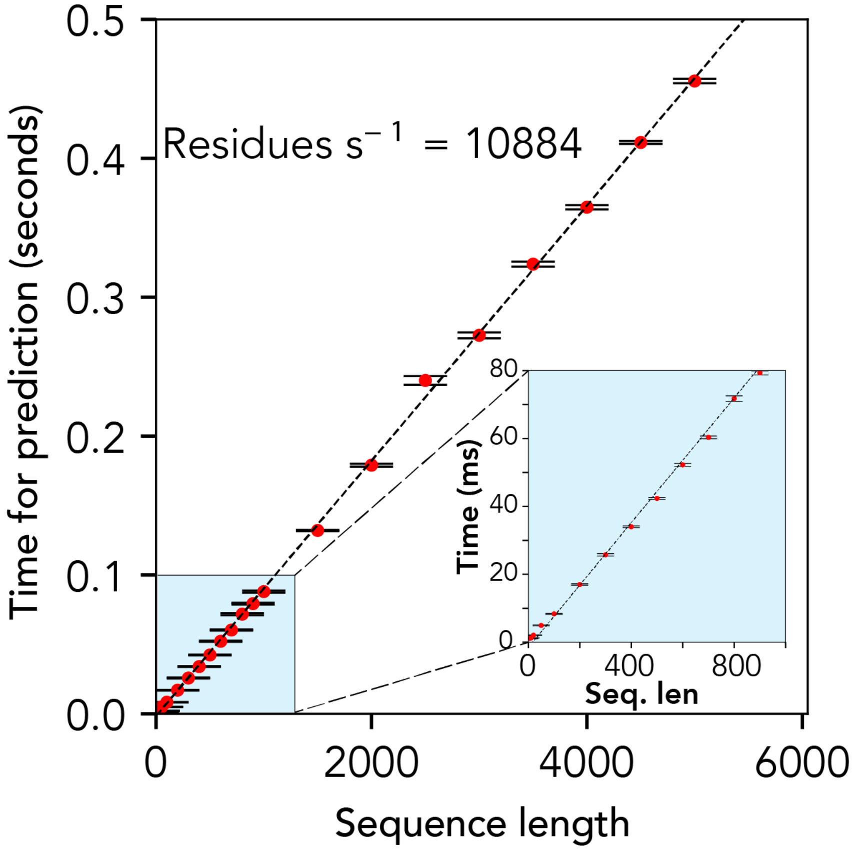
Supplemental Figure 2. Assessing the impact of disorder cutoff values on binary order and disorder classification of the Disprot Dataset (CAID). (A) F1-scores for binary predictions of disorder by metapredict compared to binary classifications of disorder from the Disprot dataset. **(B)** F1-scores for 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 3. 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 4. 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 5. 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.