Computational and Structural Biotechnology Journal 23 (2024) 1968–1977

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Taxonomy-specific assessment of intrinsic disorder predictions at residue and region levels in higher eukaryotes, protists, archaea, bacteria and viruses

Sushmita Basu, Lukasz Kurgan *

Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Keywords:
Intrinsic Disorder
Disorder Prediction
Assessment
Disordered Regions
Taxonomy
Eukaryotes
Protists
Archaea
Bacteria
Viruses

ABSTRACT

Intrinsic disorder predictors were evaluated in several studies including the two large CAID experiments. However, these studies are biased towards eukaryotic proteins and focus primarily on the residue-level predictions. We provide first-of-its-kind assessment that comprehensively covers the taxonomy and evaluates predictions at the residue and disordered region levels. We curate a benchmark dataset that uniformly covers eukaryotic, archaeal, bacterial, and viral proteins. We find that predictive performance differs substantially across taxonomy, where viruses are predicted most accurately, followed by protists and higher eukaryotes, while bacterial and archaeal proteins suffer lower levels of accuracy. These trends are consistent across predictors. We also find that current tools, except for flDPnn, struggle with reproducing native distributions of the numbers and sizes of the disordered regions. Moreover, analysis of two variants of disorder predictions derived from the AlphaFold2 predicted structures reveals that they produce accurate residue-level propensities for archaea, bacteria and protists. However, they underperform for higher eukaryotes and generally struggle to accurately identify disordered regions. Our results motivate development of new predictors that target bacteria and archaea and which produce accurate results at both residue and region levels. We also stress the need to include the region-level assessments in future assessments.

1. Introduction

Intrinsically disordered proteins (IDPs) contain one or multiple intrinsically disordered regions (IDRs), which are defined as sequence segments that lack stable structure under physiological conditions [1,2]. IDPs can be fully disordered, in which case the IDR covers the entire sequence. While IDPs can be found across all domains of life, several bioinformatics studies suggest that they are more abundant in eukaryotic proteomes [3–5]. Functionally, IDPs complement ordered/structured proteins, contributing to numerous cellular activities that include cell cycle regulation, signal transduction, transcription, post-translational modifications, and phase separation [6–8]. Given their functional importance, mis-regulation of IDPs was shown to lead to several human diseases [9–12]. Moreover, IDPs garner increasing amount of attention as potential drug targets [13–17]. Two databases, DisProt [18] and MobiDB [19], provide access to experimentally characterized IDRs, where the smaller in scale DisProt includes functionally annotated IDRs. Combining their data together results in dozens of thousands of IDPs, while the current version 2023_05 of UniProt contains around 250 million proteins sequences [20]. This large and growing IDP annotation gap motivates development of computational methods that predict intrinsically disordered residues in sequences of the millions of proteins that lack this annotation.

Well over 100 sequence-based disorder predictors were released so far [21–29]. They were comparatively evaluated in a number of community-organized assessments, starting with the fifth Critical Assessment of protein Structure Prediction (CASP5) in 2002, when six disorder predictors participated [30]. The disorder predictors were continually evaluated at the CASP events, until CASP10 in 2012 that covered 28 methods for the disorder prediction [31]. More recently, these assessments are organized and run by the intrinsic disorder prediction community. The first Critical Assessment of protein Intrinsic Disorder prediction (CAID1) was completed in 2018 and involved 32 methods [32]. CAID2 included 46 predictors and was done in 2022 [33]. These assessments are arguably more objective and impactful than other comparative studies that were done in the meantime by authors of

* Corresponding author.
E-mail address: lkurgan@vcu.edu (L. Kurgan).

https://doi.org/10.1016/j.csbj.2024.04.059
Received 5 February 2024; Received in revised form 23 April 2024; Accepted 24 April 2024
Available online 27 April 2024

individual predictors [34–39]. This is because they are run by assessors who do not participate in the evaluation, rely on relatively large and blind datasets (test proteins are not available to the authors of the participating methods before the event), apply community-screened evaluation criteria, and provide access to the datasets, ground truth annotations and predictions. These assessments provide invaluable insights concerning predictive quality, availability of predictors, and progress and growth of this field. However, the datasets used in the recent CAID1 and CAID2 events are biased towards certain parts of the taxonomy. In CAID1, 82 % of the benchmark proteins were from eukaryotes, 12 % from bacteria, 5 % from viruses and 1 % from archaea [32]. Similarly, CAID2′s datasets includes 80 % of eukaryotic proteins, 10 % bacterial, 10 % viral, and no archaeal proteins [33]. Correspondingly, the CAID results essentially reflect predictive performance on the eukaryotic proteins. This taxonomic breakdown is very different from the data in the main protein repository, UniProt [20], where 30 % of proteins are from eukaryotes, 65 % from bacteria, 2 % from viruses and 3 % from archaea. This demonstrates high levels of interest in bacterial species while disorder prediction assessments are biased towards eukaryotes. Moreover, none of the current comparative studies evaluate quality of the disorder predictions for specific parts of the taxonomy, while research shows that abundance and certain functional characteristics of disorder differ substantially across kingdoms of life [3–5,40–45].

To this end, we present first-of-its-kind taxonomy-specific assessment of disorder predictors on eukaryotic, archaeal, bacterial, and viral proteins. We rely on the CAID1 and CAID2 results to select several accurate disorder predictors. We also include two variants of disorder predictions that are derived from the protein structure generated by AlphaFold2 (AF2) [46]. We curate a test dataset that includes equal number of proteins for different parts of the taxonomy and where these sequences share below 25 % similarity with the training sequences of the selected methods. The latter ensures that this dataset is equally challenging for each tool, and simulates a scenario when predicting proteins that are dissimilar to the IDPs used for training. Importantly, we evaluate multiple aspects of the disorder predictions. Similar to the past comparative studies, we evaluate predictions at the residue level. Moreover, we study quality of the IDR predictions, by qualifying the degree of the overlap between the predicted and the native IDRs, and assessing the number and length of the predicted IDRs. Altogether, we assess several characteristics of disorder predictions across the entire taxonomic spectrum for a collection of representative methods.

2. Materials and methods

2.1. Selection of predictors

We select a collection of accurate disorder predictors using the results from CAID1 [32] and CAID2 [33]. We use the popular AUC (area under the receiver operating characteristic (ROC) curve) metric to select the ten best methods in CAID1 (on the DisProt dataset) and in CAID2 (on the Disorder-NOX dataset). In case there are multiple methods that were developed by the same research group, we select one of them that has the highest AUC score. We also remove methods which were not published in a peer-reviewed journal. These filters resulted in the removal of flDPlr, SPOT-Disorder-Single [47] and AUCpred-np [48] from the CAID1 list and flDPnn2, flDPlr and flDPlr2 from the CAID2 list. We focus on methods that consistently participated in CAID1 and CAID2, resulting in a list of five tools: flDPnn [49], EspritzD [50], rawMSA [51], Disomine [52] and SPOT-Disorder2 [53]. Given the rather large size of our test dataset that samples the entire taxonomy, we exclude SPOT-Disorder2 that has an average per-protein runtime of about 50 mins. Finally, we select the remaining four methods. These methods secure high AUC values, averaged over the two CAID assessments, and they include flDPnn, EspritzD, rawMSA, and Disomine with the average AUCs of 0.824, 0.788, 0.782, and 0.781, respectively.

Besides these four accurate predictors of disordered residues, we also include disorder predictions derived from the AF2 results. Motivated by recent studies [54–56], we apply two approaches to compute the disorder propensities from the AF2 predicted protein structures. The first approach was defined in the AF2 article [46] by using the per-residue confidence measure, predicted local-distance difference test (pLDDT), to calculate the disorder propensity. The pLDDT scores range between 0 and 100, where a higher score denotes higher reliability of the AF2′s prediction. Accordingly, we calculate the disorder propensity as (1 – pLDDT/100) and we name this prediction AF2_pLDDT. The second approach, AF2_RSA [54], uses relative solvent accessibility (RSA) scores that are calculated with DSSP [57] from the AF2 predicted protein structures, and averages these scores over a sliding window of size 25. We obtain the AF2 predicted structures from the AlphaFold Protein Structure Database (AlphaFoldDB) [58]. We note that AlphaFoldDB explicitly excludes viral proteins, and the AF2 authors made this decision for “technical reasons” [59]. This effectively means that we cannot assess AF2′s predictions for the viral proteins, while this is possible for the disorder predictors.

2.2. Dataset curation

We rely on the MobiDB database [60,61], the largest repository of disordered proteins with their structural and functional annotations obtained from computational predictions and experimentally verified sources, which are primarily PDB [62] and DisProt [18]. We collect sequences with experimentally-derived disorder annotations, which produced 22,357 proteins. We remove 18 peptides that have sequences shorter than 30 residues. Next, we remove sequences that share over 25 % similarity to the training datasets of the four selected accurate disorder predictors. This aims to make the test dataset equally difficult for each evaluated tool, and simulates predictions for proteins that are dissimilar to the training IDPs, i.e., proteins that represent a broad collection of sequences that lack disorder annotations. To do that, we cluster the 22,357 proteins with the 9721 training proteins using CD-Hit at 25 % similarity threshold and 80 % coverage [63], and we exclude all clusters that include training proteins. Correspondingly, we keep the remaining 15,221 sequences. Next, we exclude proteins that use PDB-derived annotations that rely on structures of complexes. This is because such interactions can potentially induce disorder-to-order transitions for the binding regions, which would incorrectly annotate these binding IDRs as structured. However, we keep the PDB-derived annotations where they rely on the structures of protein monomers with no ligands. Next, we obtain taxonomic details of these sequences using UniProt [20] and segregate them into five broad taxonomic groups: 1) higher eukaryotes that cover animals, plants, and fungi (1199 proteins); 2) protists that cover other eukaryotes except animals, plants, and fungi (113 proteins); 3) archaea (93 proteins); 4) bacteria (1217 proteins); and 5) viruses (124 proteins). Finally, we balance the dataset taxonomically by keeping the entire collection of the 93 proteins from archaea and randomly sampling 93 proteins from each of the other four groups, where we stratify sampling for eukaryotes to retain the original breakdown between animals, plants and fungi. We investigate whether this sampling affects the underlying characteristics of the data by comparing amino acid-level propensities for disorder between a complete dataset and the corresponding sampled dataset for each taxonomic group (Supplementary Fig. S1). As expected, since this is random sampling, the patterns of disorder enrichment are consistent between the complete and the sampled datasets where significantly enriched and depleted amino acids maintain the same bias.

We use the resulting dataset of 465 sequences that combines the five samples sets to comparatively evaluate results of the three accurate disorder predictors across the five taxonomic groups. This dataset is comparable in size to the datasets used in CASP10 (94 proteins), CAID1 (646 proteins), and CAID2 (348 proteins). We provide this dataset, which includes the UniProt accession numbers, taxonomic classification,

sequences and annotations of disordered residues/regions, in the Supplement. We also give distribution of the disorder content (fraction of disordered residues) for the proteins across the five taxonomic groups in Supplementary Fig. S2. Interestingly, these distributions that rely on the experimentally annotated disorder agree with past bioinformatics studies that estimated disorder content based on predictions, showing that eukaryotic and viral species have substantially more disorder compared to the archaea and bacteria [3,4,64,65].

2.3. Assessment metrics

Disorder predictors, including the four selected methods and the two variants of the AlphaFold2-based results, produce a numeric propensity for intrinsic disorder for each amino acid in the input sequence. These propensities are used to derive putative structural state (intrinsically disordered vs. structured) using a threshold, where amino acids with the disorder propensities $\geq$ threshold are labelled '1′ (disordered), and otherwise they are labeled '0′ (structured). Disordered residues form IDRs in the sequence, where the experimental annotations of disorder in the source databases, MobiDB and DisProt, assume that IDRs must be at least 10 consecutive residues in length [18,19]. The putative disorder is supposed to mimic this annotation and similarly generate disordered sequence segments. Correspondingly, we assess disorder predictions at the residue level and the region level, where the latter examines an overlap between the predicted and native IDRs, and compares numbers and sizes of the native and predicted IDRs.

The residue-level evaluation follows past comparative assessments and considers both propensities and binary states. In particular, we apply the popular AUC metric to evaluate the putative propensities [31–34,36,37,39]. The underlying ROC curve plots true positive rates (TPR = TP/(TP+FN)) vs. false positive rates (FPR = FP/(FP+TN)) using every unique propensity value as the threshold, where TP, TN, FN and FP are the numbers of true positives (correctly predicted disordered residues), true negatives (correctly predicted structured residues), false negatives (disordered residues incorrectly predicted as structured), and false positives (structured residues incorrectly predicted as disordered), respectively. The AUC values range between 0 (all incorrect predictions) and 1 (all predictions are correct), where 0.5 denotes random predictions. In practice, AUC scores are expected to range between 0.5 and 1. We binarize propensities to derive the putative structural state using a threshold that produces the correct disorder content (fraction of disordered residues) over the entire dataset. This calibrates predictions across different methods, allowing us to directly compare them. We assess these binary predictions using the following two metrics that were applied in the CAID and CASP experiments [31–33]:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

$$Matthews\ Correlation\ Coefficient(MCC) = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The F1 is a harmonic mean of precision = TP/(TP+FP) and TPR (also called recall), where higher values indicate better predictive performance. MCC ranges between −1 and 1 where 0 denotes predictions at random levels and higher positive value corresponds to a stronger agreement between predictions and native values; negative values are uncommon and would suggest that predicted state is opposite to the native state.

We also evaluate quality of the predicted IDRs. While this aspect was not assessed in the CASP and CAID experiments, similar evaluations were done for the secondary structure and transmembrane region predictions, which also form segments in the sequence [66–70]. We generate the binary state with the thresholds that generate correct disorder content over the entire dataset when setting the minimum IDR length to 10 residues, which is consistent with the minimal regions sizes in the source DisProt and MobiDB databases [18,19]. We compare the number of putative IDRs for each of the unique IDR lengths, with the corresponding native IDR counts using mean absolute error (MAE). For a given set of n unique IDR lengths with native IDR counts $a_1, a_2, ..., a_n$ and the predicted IDR counts $x_1, x_2, ..., x_n$, the MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - a_i|$$

This metric computes an average count by which number of predicted IDRs differ from number of native IDRs across all native lengths of IDR. For example, MAE of 20 means that on average (across different IDR sizes) the number of predicted IDRs differ by 20 from the number of native IDRs. We also compare distributions of native and predicted IDRs (i.e., plots of the numbers of IDRs across different IDR sizes) using the Kolmogorov-Smirnov test, and we apply the p-value that this test generates as the metric. This metric measures the difference in shape between the distributions of the predicted and the native IDRs, where higher p-values correspond to higher likelihood that the native and predicted IDR distributions are similar. Besides the number and sizes of predicted IDRs, we assess the degree of overlap between the native and predicted IDRs using the segment overlap (SOV) metric [71], which was originally developed for the assessment of the secondary structure regions and was used to assess disorder predictions [38,72]. SOV complements the MAE and p-value based evaluation since it considers position of the predicted IDRs in the sequence relative to the position of the native IDRs. SOV ranges between 0 (no overlap) and 1 (perfect overlap) and is calculated per protein. Moreover, since sequence length varies across taxonomy, we compute weighted average of these values over the corresponding proteins in a given taxonomic group where weights correspond to the protein length. This facilitates direct comparison of the SOV values across taxonomy.

2.4. Statistical tests

We quantify statistical significance of differences when comparing predictions across taxonomy. We compare results generated for diverse sub-sampled collections of proteins from the same taxonomic groups. Specifically, we randomly sample 10 sets of 50 % proteins from a given taxonomic group and compare with the corresponding 10 sampled results from another group. We use the student t-test when the corresponding data (i.e., measured AUC, F1, MCC, MAE, p-value and SOV values) are normal, and otherwise we apply the Wilcoxon rank-sum test. We determine normality using the Anderson-Darling test at 0.05 significance.

3. Results

3.1. Residue-level assessment

We compare quality of the residue-level predictions produced by the four accurate disorder prediction methods (EspritzD, flDPnn, rawMSA and Disomine) on balanced collections of proteins from five diverse taxonomic groups (higher eukaryotes, protists, archaea, bacteria and viruses) in Table 1. The confusion matrices and additional metrics that assess binary predictions are in Supplementary Table S1.

Table 1 reveals that the four disorder predictors perform reasonably well with AUCs ranging from moderate (0.69 for Disomine for bacteria) to high (0.87 for EspritzD and Disomine for viruses). This stems from the fact that we sampled arguably currently the most accurate disorder predictors. Importantly, predictive performance varies widely across taxonomy, with substantial and statistically significant differences between the best and the worst performing taxonomic groups, i.e., AUC of 0.87 vs. 0.70 for EspritzD (p-value$\leq$0.01); 0.86 vs. 0.71 for flDPnn (p-value$\leq$0.01), 0.85 vs. 0.71 for rawMSA (p-value$\leq$0.01) and 0.87 vs 0.69

Table 1
Residue-level assessment of predictive performance for the six predictors and the five diverse taxonomic groups. We sort the taxonomic groups in the descending order of values for a given performance metric. We report medians of the metrics that we calculate over the 10 sampled datasets (see "Statistical test" section for details). We summarize results of the statistical significance analysis in the *x/y* format next to the reported median value where *x* and *y* compare against the best and worst predicted taxonomic group, respectively, and where ** and * denote statistically significant differences with *p*-values ≤ 0.01 and ≤ 0.05, respectively, while = denotes differences that are not statistically significant (*p*-value > 0.05). Predictions from AF2_pLDDT and AF2_RSA results are unavailable (UA) for the viral proteins.

Methods	AUC		F1		MCC	
	Taxonomic groups	Score	Taxonomic groups	Score	Taxonomic groups	Score
EspritzD	Viruses	0.866 /* *	Viruses	0.438 /* *	Viruses	0.388 /* *
	Protists	0.753 * */=	Protists	0.298 * */* *	Protists	0.213 * */* *
	Higher eukaryotes	0.738 * */=	Higher eukaryotes	0.274 * */* *	Higher eukaryotes	0.164 * */=
	Archaea	0.725 * */=	Archaea	0.246 * */*	Archaea	0.151 * */=
	Bacteria	0.696 * */	Bacteria	0.184 * */	Bacteria	0.137 * */
flDPnn	Viruses	0.856 /* *	Viruses	0.481 /* *	Viruses	0.419 /* *
	Higher eukaryotes	0.824 * /* *	Higher eukaryotes	0.466 = /* *	Higher eukaryotes	0.396 = /* *
	Protists	0.797 * */* *	Protists	0.371 * /* *	Protists	0.302 * /* *
	Archaea	0.756 * */=	Archaea	0.317 * */=	Archaea	0.262 * */=
	Bacteria	0.706 * */	Bacteria	0.303 * */	Bacteria	0.241 * */
rawMSA	Viruses	0.853 /* *	Viruses	0.378 /* *	Viruses	0.316 /* *
	Protists	0.815 * */* *	Protists	0.375 = /* *	Protists	0.296 = /* *
	Higher eukaryotes	0.785 * */=	Higher eukaryotes	0.351 = /* *	Higher eukaryotes	0.262 * */=
	Archaea	0.772 * */=	Bacteria	0.314 = /* *	Bacteria	0.260 * /*
	Bacteria	0.713 * */	Archaea	0.224 * */	Archaea	0.193 * */
Disomine	Viruses	0.866 /* *	Viruses	0.374 /* *	Viruses	0.315 /* *
	Protists	0.765 * */* *	Protists	0.353 = /* *	Protists	0.301 = /* *
	Higher eukaryotes	0.732 * */*	Higher eukaryotes	0.306 * /* *	Higher eukaryotes	0.208 * */*
	Archaea	0.724 * */=	Archaea	0.245 * */=	Archaea	0.179 * */* *
	Bacteria	0.688 * */	Bacteria	0.218 * */	Bacteria	0.148 * */
AF2_pLDDT	Archaea	0.818 /* *	Protists	0.277 /=	Archaea	0.237 /*
	Protists	0.817 = /* *	Bacteria	0.268 = /*	Bacteria	0.222 = /*
	Bacteria	0.813 = /* *	Higher eukaryotes	0.267 = /*	Protists	0.167 * /=
	Higher eukaryotes	0.736 * */	Archaea	0.221 = /	Higher eukaryotes	0.156 * /
	Viruses	UA	Viruses	UA	Viruses	UA
AF2_RSA	Archaea	0.845 /* *	Bacteria	0.307 /* *	Archaea	0.265 /* *
	Bacteria	0.823 = /* *	Protists	0.280 = /=	Bacteria	0.258 = /* *
	Protists	0.798 * */* *	Higher eukaryotes	0.228 * /=	Protists	0.182 * */=
	Higher eukaryotes	0.724 * */	Archaea	0.221 * */	Higher eukaryotes	0.101 * */
	Viruses	UA	Viruses	UA	Viruses	UA

for Disomine (*p*-value≤0.01). Moreover, we note a consistent sorted order of taxonomic groups, with the most accurate predictions for eukaryotes and viruses, followed by archaea, and the least accurate results for bacteria. This trend is similar across different predictors and metrics. The four tools produce equally very accurate predictions for the viral proteins, with AUCs ranging between 0.85 (rawMSA) and 0.87 (EspritzD and Disomine). On the other end of the spectrum, the lowest performance is for the bacterial proteins, with modest and similar AUCs around 0.70. We observe a more variable levels of performance for the higher eukaryotes and protists, with flDPnn securing high AUCs of 0.82 and 0.80, respectively, and Disomine obtaining substantially lower AUCs of 0.73 and 0.77, respectively. These higher levels of performance for flDPnn explain why this tool performed better than the other three methods in the CAID assessments where large majority of the test proteins are from eukaryotes [32,33]. Finally, predictions for archaea are characterized by modest levels of performance, with AUCs between 0.72 (Disomine) and 0.77 (rawMSA). Moreover, AUCs for archaea are not statistically better than AUCs for bacteria (*p*-value > 0.05), suggesting that these two taxonomic groups suffer similarly low levels of predictive quality. Analysis based on the F1 and MCC metrics for the binary state predictions leads to consistent observations. The highest correlations are for viral proteins (MCC of 0.42 for flDPnn), higher eukaryotes (MCC of 0.40 for flDPnn), and protists (MCC of 0.30 for flDPnn, rawMSA and Disomine). We find substantially lower MCCs for archaea and bacteria, with the most accurate results for flDPnn (MCC of 0.26 for archaea) and rawMSA (MCC of 0.26 for bacteria). Altogether, we observe that the residue-level predictive performance of the four disorder predictors varies broadly across the taxonomy, with viruses and eukaryotes predicted accurately, and archaea and bacteria securing much lower and modest levels of performance.

We also assess AF2 derived disorder predictions, AF2_pLDDT and AF2_RSA, in the four taxonomic groups which exclude viruses for which AF2 does not provide predictions. In contrast to the disorder predictors, both AF2 variants accurately predict disorder propensities for archaeal proteins, with AUCs of 0.85 and 0.82 for AF2_RSA and AF2_pLDDT, respectively, compared to the best disorder predictor that secures AUC of 0.77. They also produce similarly accurate results for protists and bacteria while performing at modest levels of performance for the higher eukaryotes, i.e., AUC = 0.74 for AF2_pLDDT and AUC = 0.72 for AF2_RSA when compared to AUC of 0.82 for flDPnn. The relatively low levels of performance for the higher eukaryotes may explain why the AF2_pLDDT and AF2_RSA predictors were outperformed by several disorder predictors on the Disorder-NOX dataset in CAID2, which is composed largely of eukaryotic proteins [33]. Moreover, the high AUC values for archaea and protists are coupled with disproportionally lower quality of the binary state predictions. In particular, the F1 scores of 0.22 for archaea and 0.28 for protist for the best variant of AF2 are much lower than F1 of 0.32 by flDPnn and F1 of 0.37 by flDPnn and rawMSA, respectively. This suggests that while AF2 derived predictions generate accurate residue-level propensities, they might not perform as well when predicting IDRs based on the putative binary disorder, which we evaluate in the next section.

3.2. Region-level assessment

We assess two complementary aspects of the region-level predictive performance: ability to mimic the native distribution of IDRs of different sizes and the degree of overlap between putative and native IDRs. In Fig. 1, we show the cumulative counts of IDRs across different IDR length values, separately for each taxonomic group. We find that the length and numbers of the native IDRs (red plots in Fig. 1) differ across taxonomy, in spite of the fact that the number of proteins is the same.

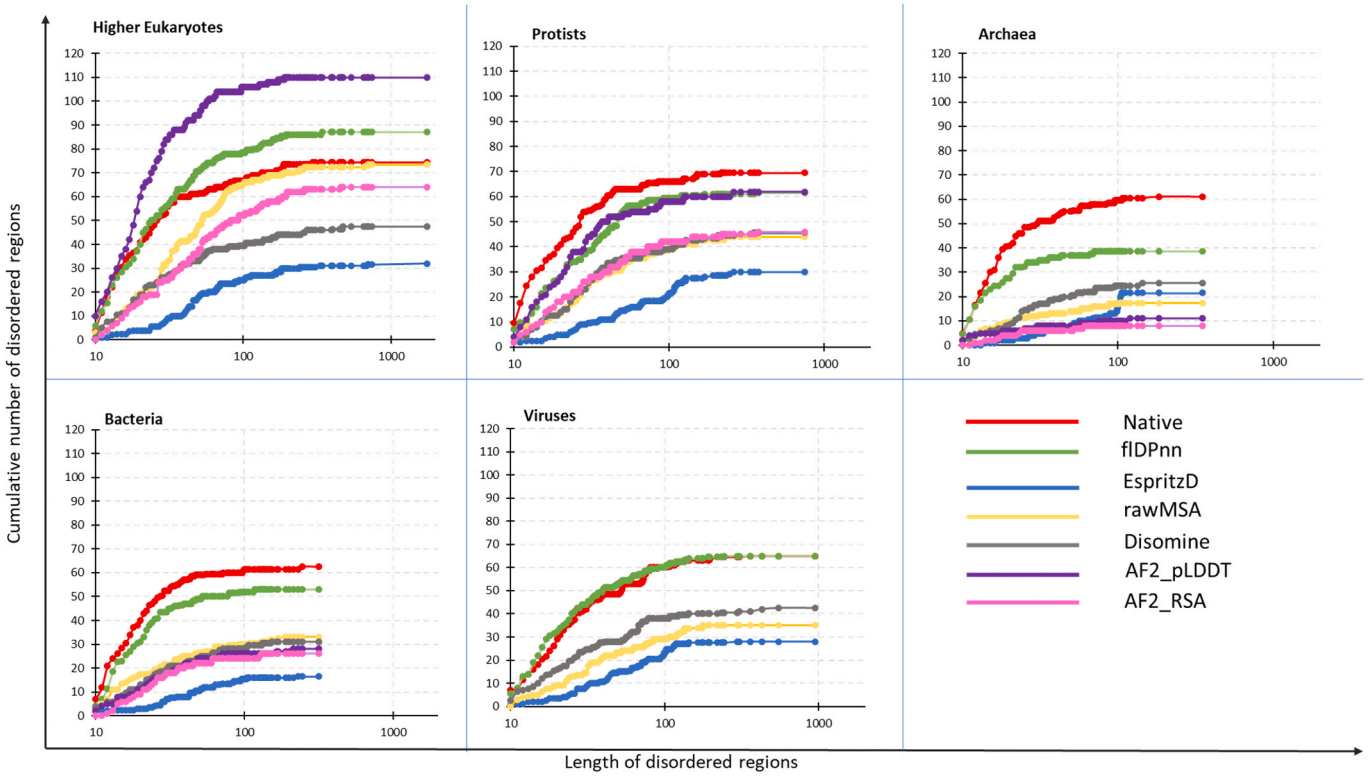

Fig. 1. Cumulative distributions of native IDR sizes (red plots; based on the disorder annotations from MobiDB) and IDR sizes predicted by the four selected and accurate disorder predictors (blue for EspritzD; green for flDPnn; yellow for rawMSA; grey for Disomine) and two AF2 derived disorder predictors (purple for AF2_pLDDT; pink for AF2_RSA) across the diverse taxonomic groups. Results from AF2_pLDDT and AF2_RSA results are unavailable for the viral proteins.

Higher eukaryotes have the highest number and the longest IDRs, with some regions longer than 1000 amino acids. Some IDRs in viral proteins are also relatively long, while IDRs in the archaeal, bacterial and protist proteins are relatively short and there are fewer of them in the latter two taxonomic groups. These observations align with the distributions of the disorder content (Supplementary Fig. S1), which show that eukaryotic and viral proteins have on average higher disorder content, with some proteins having more than 80 % disordered residues.

Fig. 1 compares distributions of the numbers and sizes of native IDRs with the putative IDRs generated by the four disorder predictors for each of the five taxonomic groups. We find that while some tools produce distributions that are relatively similar to the experimental data, others predict fewer and longer IDRs than expected. The flDPnn's plots (green lines in Fig. 1) match relatively well with the experimental IDRs (red lines in Fig. 1) for higher eukaryotes, protists, bacteria and viruses. The rawMSA method under-predicts the number of short IDRs while generating relatively correct numbers of longer IDRs (yellow lines in Fig. 1). EspritzD under-predicts the number of short and moderately long IDRs and over-predicts long IDRs (blue lines in Fig. 1). Disomine similarly under predicts the short IDRs in all taxonomic groups, however, it predicts the amounts of longer IDRs relatively correct, particularly in higher eukaryotes, protists and viruses (grey lines in Fig. 1). Overall, all methods under-predict the number of IDRs for protists, archaea, bacteria, although flDPnn makes smaller mistakes when compared to rawMSA, EspritzD, and Disomine.

Table 2 quantifies differences between the native distributions and each of the four predicted distributions from Fig. 1 using the mean absolute error (MAE) and *p*-value. MAE is an average difference in the number of IDRs across different region sizes. The *p*-value assesses significance of the differences in the shape of the distributions, where $p > 0.05$ means that they are not statistically different. We note that each taxonomic group has a unique range of errors that is defined by its total count of native IDRs. Thus, we normalize the MAE values for the taxonomic groups to the widest range of error, observed in eukaryotes, which is 0 to 75. This allows us to directly compare the MAE values between different taxonomic groups. EspritzD (blue plots in Fig. 1) has high MAE scores for all taxonomic groups (Table 2), which implies that number of IDRs predicted by this method is very different from the native IDR counts (red plot in Fig. 1). Its lowest/best MAE of 41 is for viral proteins, which still exceeds half of the total number of native IDRs. The putative IDRs produced by EspritzD differ from the native IDRs not only in their number, but also in the shape of the IDR distribution. The low *p*-values in Table 2 reveal that the distributions of the putative IDRs generated by EspritzD are significantly different from the distributions of experimental IDRs (p-value < 0.01). Disomine (grey lines in Fig. 1) secures similar MAE scores of around 25, for viruses, protists and higher eukaryotes, while its errors are significantly higher for archaea and bacteria (p-values < 0.01). The shape of predicted IDR distributions for viruses and higher eukaryotes are similar to the respective native distributions (p-values > 0.13), whereas they are significantly different for protists and archaea (p-values < 0.01). The rawMSA method (yellow plots in Fig. 1) secures a relatively low MAE of 11 for higher eukaryotes but its MAEs are high for the other taxonomic groups (Table 2), suggesting that it substantially under-predicts the number of IDRs for the viral, bacterial and archaeal proteins. The shape of the distributions generated by rawMSA are similar to the native distributions for higher eukaryotes and protists (p-values ≥ 0.17), while being significantly different for archaea (p-value < 0.01). The MAE scores of flDPnn (green plots in Fig. 1) are low for eukaryotes, viruses and bacteria (Table 2). The only lower quality result is for archaea where flDPnn's MAE is 19. Moreover, the distributions of IDRs generated by flDPnn share similar shapes with the distributions of native IDRs across the five taxonomic groups (p-values≥ 0.10). To summarize, we find that flDPnn produces relatively accurate numbers and sizes of IDRs while EspritzD underpredicts short regions and produces significantly different distributions of region sizes when compared to the experimental data. Disomine

Table 2
Region-level assessment of the distributions of the numbers and sizes of putative IDRs generated by the six predictors for the five diverse taxonomic groups. We quantify the mean absolute error (MAE) between the number of native and predicted IDRs over different lengths of the regions, and *p*-value that evaluates significance of differences in the shapes of the resulting plots (see "Assessment Metrics" section for details). We sort the taxonomic groups from the best to the worst performance, i.e., in the ascending order by their MAE values and the descending order by their *p*-values. We report medians of the metric that we calculate over the 10 sampled datasets (see "Statistical test" section for details). We summarize results of the statistical significance analysis in the *x*/*y* format next to the reported median value where *x* and *y* compare against the best and worst predicted taxonomic group, respectively, and where * * and * denote statistically significant differences with *p*-values ≤ 0.01 and ≤ 0.05, respectively, while = denotes differences that are not statistically significant (*p*-value > 0.05). Predictions from AF2_pLDDT and AF2_RSA results are unavailable (UA) for the viral proteins.

Methods	Taxonomic groups	MAE	Taxonomic groups	*p*-value (distribution shape)
EspritzD	Viruses	41.32 /* *	Viruses	1.09E-03 /* *
	Higher eukaryotes	41.52 = / * *	Higher eukaryotes	6.21E-04 = /* *
	Protists	44.24 = / * *	Protists	2.72E-05 * */=
	Bacteria	52.64 * */ =	Archaea	7.80E-07 * */=
	Archaea	52.77 * */	Bacteria	4.05E-07 * */
flDPnn	Protists	6.31 /* *	Viruses	0.648 /* *
	Viruses	6.48 = / * *	Protists	0.618 = /* *
	Higher eukaryotes	7.83 = / * *	Higher eukaryotes	0.441 = /* *
	Bacteria	9.94 = / * *	Bacteria	0.312 = /* *
	Archaea	19.76 * */	Archaea	0.098 * */
rawMSA	Higher eukaryotes	10.48 /* *	Higher eukaryotes	0.451 /* *
	Protists	29.39 * */ * *	Protists	0.174 * */* *
	Viruses	32.14 * */ * *	Viruses	0.066 * */*
	Bacteria	34.51 * */ * *	Bacteria	0.063 * */*
	Archaea	46.75 * */	Archaea	3.93E-04 * */
Disomine	Viruses	24.72 /* *	Viruses	0.472 /* *
	Protists	24.87 = / * *	Higher eukaryotes	0.135 * */* *
	Higher eukaryotes	26.82 = /*	Bacteria	0.016 * */=
	Bacteria	37.08 * */ =	Protists	0.004 * */=
	Archaea	40.56 * */	Archaea	0.003 * */
AF2_pLDDT	Protists	8.27 /* *	Protists	5.66E-15 /* *
	Higher eukaryotes	37.33 * */ * *	Archaea	4.15E-42 * /*
	Bacteria	39.21 * */ * *	Bacteria	5.86E-48 * */=
	Archaea	53.44 * */	Higher Eukaryotes	5.66E-76 * */
	Viruses	UA	Viruses	UA
AF2_RSA	Higher eukaryotes	15.18 /* *	Higher Eukaryotes	3.44E-26 /* *
	Protists	19.64 * / * *	Protists	6.15E-42 = /* *
	Bacteria	41.71 * */ * *	Archaea	5.34E-44 = /* *
	Archaea	56.10 * */	Bacteria	5.71E-48 * /
	Viruses	UA	Viruses	UA

underpredicts the short and moderately long IDRs, but it accurately mimics the shape of the native IDR distributions for viruses and higher eukaryotes that are enriched in long IDRs. The rawMSA tool provides reasonably accurate results for higher eukaryotes while underpredicting IDRs and in particular short IDRs for the other four taxonomic groups.

Table 3 reports the region-level SOV scores that quantify the degree of overlap between the native and putative IDRs in protein sequences. We find that SOV values vary by a factor of 2 across taxonomic groups, between 0.38 (EspritzD for archaea) and 0.83 (Disomine for viruses). Interestingly, the four disorder predictors consistently produce high SOV values, over 0.77, for the viral proteins. EspritzD generates high SOV of 0.66 for the higher eukaryotes, followed by Disomine with SOV of 0.61, while the other two methods secure lower scores at around 0.55 for these proteins. Results for protists and bacteria are also relatively accurate and consistent across the four disorder predictors, with SOV of rawMSA at 0.52 for protists and SOV of EspritzD, flDPnn and Disomine at 0.53 for bacteria. However, we note relatively poor quality of the region-level overlap for archaea, with SOV of 0.40 for the best flDPnn and rawMSA. Altogether, similar to the residue-level evaluation, we find that the segment overlap differs across taxonomy where viruses are predicted very accurately, eukaryotes and bacteria are predicted with modest performance, and archaea suffers relatively low predictive quality. Importantly, these trends are consistent across the four disorder predictors.

We also perform the region-level assessment for the two AF2 derived predictions for protists, higher eukaryotes, archaea and bacteria, excluding viruses where AF2 does not generate predictions. AF2_pLDDT (purple line in Fig. 1) predicts the size and number of IDRs relatively well for protists, substantially overpredicts short and moderately long IDRs in higher eukaryotes, and severely underpredicts IDRs in archaea

Table 3
Region-level assessment of the segment overlap (SOV) between the native IDRs and the putative IDRs generated by the six predictors for the five diverse taxonomic groups. Higher SOV scores indicate larger degree of the overlap. We sort the taxonomic groups in the descending order of SOV values. We report median SOVs that we calculate over the 10 sampled datasets (see "Statistical test" section for details). We summarize results of the statistical significance analysis in the *x*/*y* format next to the reported median value where *x* and *y* compare against the best and worst predicted taxonomic group, respectively, and where * * and * denote statistically significant differences with *p*-values ≤ 0.01 and ≤ 0.05, respectively, while = denotes differences that are not statistically significant (*p*-value > 0.05). Predictions from AF2_pLDDT and AF2_RSA results are unavailable (UA) for the viral proteins.

Methods	Taxonomic groups	SOV
EspritzD	Viruses	0.807/**
	Higher eukaryotes	0.660**/**
	Bacteria	0.525**/**
	Protists	0.502**/**
	Archaea	0.375**/
flDPnn	Viruses	0.774/**
	Higher eukaryotes	0.550**/**
	Bacteria	0.529**/**
	Protists	0.508**/**
	Archaea	0.404**/
rawMSA	Viruses	0.776/**
	Higher eukaryotes	0.574**/**
	Protists	0.521**/**
	Bacteria	0.518**/**
	Archaea	0.404**/
Disomine	Viruses	0.831/**
	Higher eukaryotes	0.614**/**
	Bacteria	0.532**/**
	Protists	0.494**/**
	Archaea	0.393**/
AF2_pLDDT	Bacteria	0.543/**
	Higher eukaryotes	0.522**/**
	Archaea	0.416**/**
	Protists	0.405**/**
	Viruses	UA
AF2_RSA	Bacteria	0.577/**
	Higher eukaryotes	0.535**/**
	Protists	0.472**/**
	Archaea	0.418**/**
	Viruses	UA

and bacteria. Correspondingly, Table 2 reveals that AF2_pLDDT secures low MAE of 8.3 for protists, and only flDPnn performs better for these proteins with MAE of 6.3. However, AF2_pLDDT's MAEs for the other three taxonomic groups are high and exceed 37. Moreover, in spite of the low MAE in protists, AF2_pLDDT fails to mimic the shape of their native IDR distribution (p-value < 0.01) and performs similarly poorly for the other three taxonomic groups (p-value < 0.01). Table 3 shows that AF2_pLDDT generates modest levels of performance when considering the overlap between its putative IDRs and the native IDRs. The SOV scores range between 0.40 and 0.54, and they are particularly low for protists, at 0.40, where rawMSA disorder predictor secures SOV of 0.52. This suggests that while AF2_pLDDT predicts the number and sizes of IDRs for protists relatively well, their location in the sequence is not predicted as well.

The AF2_RSA method (pink line in Fig. 1) underpredicts short IDRs in the four taxonomic groups but predicts longer IDRs relatively well for higher eukaryotes and protists. Table 2 shows that AF2_RSA secures reasonably low MAE of 15 for the higher eukaryotes (two disorder predictors, flDPnn and rawMSA obtain lower/better MAEs) but these errors are significantly worse for protists, bacteria and archaea (p-value < 0.05). Moreover, the shape of IDR distributions derived from the AF2_RSA's predictions is significantly different from the corresponding native distribution for the four taxonomic groups where it can produce results (p-value < 0.01). Table 3 shows that SOV scores of AF2_RSA are relatively good for bacteria and archaea but they lag behind the disorder predictors for protists and higher eukaryotes. To compare with the disorder predictions, an average SOV across the four taxonomic groups for ESpritzD is 0.52 while AF2_RSA and AF2_pLDDT obtain SOV averages of 0.50 and 0.47, respectively. Moreover, the four disorder predictors secure SOV scores between 0.77 and 0.83 for the viral proteins where AF2 does not make predictions. Altogether, the region-level assessment reveals that the AF2 derived methods are outperformed by the disorder predictors, particularly in the context of reconstructing the distributions of the IDRs sizes and the extend of the overlap between the predicted and native IDRs in sequences. This aligns well with the relatively low quality of the binary state predictions generated by AF2_pLDDT and AF2_RSA that we observe in Table 1.

4. Summary and discussion

While disorder predictors were assessed in numerous studies [31–39], these works do not consider the taxonomic diversity of the underlying test proteins. We evaluate results produced by a representative collection of four accurate disorder predictors and AF2 over the entire taxonomic spectrum including higher eukaryotes, protists, bacteria, archaea and viruses. Moreover, we analyze three diverse aspects of the predictive performance including residue-level predictions, number and sizes of the putative IDRs, and segment overlap between putative and native IDRs. Given the comprehensive nature of our analysis, i.e., five taxonomic groups, three aspects of predictive performance, and six predictors, we summarize these results using a rank-based approach focusing on the taxonomy. We could not include the AF2_pLDDT and AF2_RSA predictors in this analysis since they do not provide results for viral proteins, and so we discuss AF2 results separately in subsequent paragraphs. We rank the five taxonomic groups for each of the four disorder predictors and assessment metric and we average these ranks across the methods. Ranks of 1 and 5 correspond to the best and worst predicted taxonomic groups, respectively. Moreover, we score the taxonomic groups that secure near-random levels of performance for a particular metric with the worst/highest rank. For the residue-level assessments, the near-random performance corresponds to AUC, F1 and MCC around 0.5, 0.1 or lower, and 0.1 or lower, respectively. In the MAE case, we consider scores higher than 50 % of the maximum error range as near-random. Similarly, the p-values < 0.01 corresponds to predicted IDR distributions which are significantly different from native distributions and hence, are equivalent to near-random distributions. Finally, ref. [71] shows that SOVs below 0.19 represent overlap between two random proteins with a p-value close to one.

Fig. 2 shows the corresponding average ranking of taxonomic groups for each of the considered metrics of predictive performance. We note that lower rank means that a given taxonomic group is consistently (across different methods) predicted with higher levels of the predictive performance.

Fig. 2 reveals that predictive quality differs substantially across taxonomy, particularly for the residue-level metrics (blue bars) and the segment overlap (yellow bar). These metrics, which include AUC, F1, MCC and SOV, are in good agreement with each other, and show that viruses are consistently predicted with the highest levels of performance, followed by protists and higher eukaryotes, while bacterial and archaeal proteins are predicted relatively poorly. We note that the small standard deviations (red capped lines) signify the fact that these results are consistent across the four predictors. On the other hand, the MAE and p-value metrics (green bars), which evaluate how well the predictions mimic the distributions of the number and sizes of the native IDRs, show that ranking is inconsistent across predictors. This is because the values are more similar across the taxonomic groups and the corresponding standard deviations are also on average higher. This stems from the fact that only flDPnn provides reasonably accurate distributions for eukaryotes, viruses and bacteria, while the other predictors underpredict short IDRs, except for rawMSA that generates accurate results for higher eukaryotes. This aspect of disorder predictions is particularly inaccurate for archaeal proteins where all four methods fail to provide accurate distributions. Altogether, we find that disorder predictors perform at substantially different levels of predictive quality for different parts of taxonomy. The residue-level assessment and the SOV scores suggest that predictions are the most accurate for viral proteins (AUC of 0.87 and SOV > 0.80 for EspritzD and Disomine) and similarly very accurate for higher eukaryotes (AUC of 0.82 and SOV of 0.55 for flDPnn), and protists (AUC of 0.82 and SOV of 0.52 for rawMSA). However, the predictions for archaeal and bacterial proteins are only modestly accurate (AUC ≤ 0.77 and SOV ≤ 0.40 for archaea; AUC $\leq$ 0.71 and SOV ≤ 0.53 for bacteria). Similarly, the current disorder predictors also struggle with reproducing the distribution of the numbers and sizes of IDRs for the archaeal proteins.

We hypothesize that possible reasons for the varying performance of the disorder across taxonomy could be a taxonomic bias in their training datasets and/or differences in the compositional bias of amino acid, i.e., differences in the amino acid-level propensities for disorder that cannot be adequately addressed by the taxonomy-agnostic predictive models. We were able to collect the taxonomic details of the training data for two methods, flDPnn and rawMSA. The taxonomic breakdown of the flDPnn's training sequences is approximately 75 % eukaryotes, 16 % bacteria, 7 % viruses and 2 % archaea. Similarly, the rawMSA's training sets are composed of around 69 % eukaryotes, 20 % bacteria, 9 % viruses and 2 % archaea. While the substantial enrichment in the eukaryotic proteins may explain why these predictors perform well for the eukaryotes, the high levels of predictive quality for viruses and similar levels for bacteria and archaea, in spite of 8 to 10 times differences in their numbers in the training data, do not align with this explanation.

IDRs are known to have compositional bias at the amino acid level [73–76] and we posit that this bias might be different across taxonomy. This, in turn, could cause problems for the taxonomy-agnostic models that generate predictions from the amino acid sequences. We compute the amino acid bias of the intrinsic disorder and present these results in Supplementary Fig. S3. We were able to reproduce the bias identified in the past studies [75,76] when using the dataset of 465 proteins that combines the sampled sets of 93 proteins from the five taxonomic group (top left panel in Supplementary Fig. S3). More specifically, we find that the disorder bias is significantly negative (p-value < 0.05; green dotted box in Supplementary Fig. S3) for the order promoting residues reported in these studies. Similarly, we find that the amino acids with the

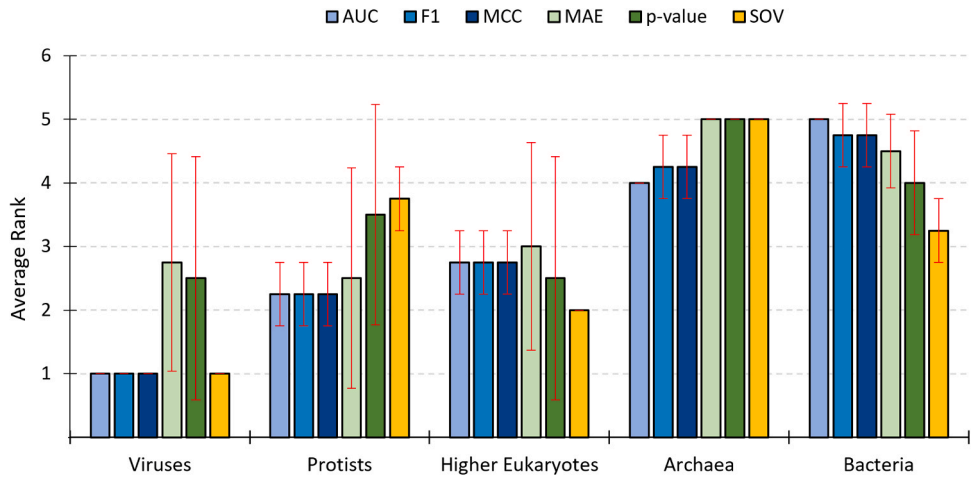

Fig. 2. Average rank (over the four disorder predictors) for the predictive performance across the five taxonomic groups. The color-coded bars identify different types of evaluations where shades of blue are for the residue-level evaluations (AUC, F1 and MCC), shades of green for the region-level distribution evaluations (MAE, and *p*-value for distribution), and yellow for the region-level segment overlap (SOV). The taxonomic groups are arranged in the ascending order of the average rank based on AUC, where a lower rank depicts higher predictive quality. The error bars (red capped lines) are the standard deviations associated with the averages.

significantly positive bias for disorder (*p*-value < 0.05; red dotted box in Supplementary Fig. S3), except for Asn, were previously reported as disorder-promoting. Moreover, out of the four amino acids that did not produce a significant bias in our analysis on the dataset with 465 proteins (Ala, Thr, Asp, and Met), three of them (Thr, Asp and Met) were reported as neutral (neither disorder nor order promoting) in ref. [76]. Our analysis of the compositional bias for the five taxonomic groups (Supplementary Fig. S3) reveals that the disorder promoting amino acids do not switch to act as the order promoting residues in any of the five taxonomic groups and vice versa. This indicates that the compositional bias across the taxonomic groups is generally consistent. However, the degree and significance of the bias varies substantially, where order and disorder promoting amino acids in some parts of the taxonomy become neutral in other taxonomic groups and some neutral amino acids may shift to be disorder or order promoting. For example, Val that has significantly negative bias for disorder (i.e., order promoting bias) in eukaryotic and viral proteins has neutral propensity for disorder in archaea and bacteria. We quantify the degree of agreement between the reference compositional bias on the entire dataset that combines all taxonomic groups and the individual taxonomic groups by counting how many amino acids match their reference disorder and order bias (red and green bars in Supplementary Fig. S3). Out of the 16 disorder and order promoting amino acids from the entire dataset, 14 matches in viruses, 13 in higher eukaryotes, 12 in protists, 10 in archaea, and 9 in bacteria. This correlates well with the varying levels of predictive performance across these taxonomic groups, where AUCs that we averaged across the four disorder predictors are 0.86, 0.79, 0.77, 0.74 and 0.70 for viruses, protists, higher eukaryotes, archaea and bacteria, respectively. The corresponding Pearson correlation coefficient between the numbers of matches and the average AUCs is 0.91. This suggests that the degree of divergence from the overall taxonomy-agnostic disorder bias may partly explain the corresponding differences in the predictive performance for the disorder predictors.

The two AF2-derived disorder predictors, AF2_pLDDT and AF2_RSA, produce accurate residue-level propensities for the archaea, bacteria and protists while being outperformed by the disorder predictors for the higher eukaryotes. However, these two predictors secure lower levels of performance for the binary state predictions (median F1 of 0.27 and median MCC of 0.20) when compared to the disorder predictors (median F1 of 0.31 and median MCC of 0.23 when excluding viruses for which AF2 does not produce predictions). Correspondingly, we find that AF2_pLDDT and AF2_RSA predicted IDRs suffer lower degree of overlap with the native IDRs (median SOV of 0.49) when contrasted with the disorder predictors (median SOV of 0.52 when excluding viruses). The AF2 derived results are also outperformed by the disorder predictors in the context of reconstructing the distributions of the IDRs sizes (median *p*-value of $2.1 * 10^{-42}$ for AF2 vs. $3.9 * 10^{-2}$ when excluding viruses; higher value is better). The lower quality when predicting IDRs can be explained by the fact that AF2 was designed to predict protein structure and so it should excel in predicting ordered residues rather than disordered residues and regions. The overall observation that AF2 derived predictions of disorder suffer lower quality than the predictions generated by state-of-the-art disorder predictors is in line with other studies [54,56].

To sum up, we assess predictive performance of four representative disorder predictors along with two AF2 derived methods on protein sequences in the five taxonomic groups. Our study sheds light on substantial limitations of the current disorder predictors along with the challenges in using AF2 for the disorder prediction. We see the need for the disorder prediction community to develop a new generation of methods that aim to provide more accurate results at the residue and the region levels for the bacterial and archaeal proteins. This need is particularly acute for the archaeal organisms where IDRs were found to be important for their adaptation to hostile habitats [45] and for which none of the tools produced accurate region-level results. Moreover, the new tools should strive to more accurately reproduce the distributions of IDR numbers and sizes since we found that only one method, flDPnn, was able to do that reasonably well. Moreover, we advocate for the inclusion of the region-level assessments in the future comparative studies, so that progress and performance on this often-neglected aspect of the disorder prediction is adequately measured.

Funding

This work was funded in part by the National Science Foundation (DBI2146027 and IIS2125218) and the Robert J. Mattauch Endowment funds to LK.

CRediT authorship contribution statement

Sushmita Basu: Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Lukasz Kurgan:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

Authors declare no conflict of interests.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.059.

References

[1] Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Zsuzsanna, et al. What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. Intrinsically Disord Proteins 2013;1(1): e24157.
[2] Oldfield CJ, Uversky VN, Dunker AK, Kurgan L. Introduction to intrinsically disordered proteins and regions. Intrinsically Disord Protein: Dyn Bind Funct 2019.
[3] Peng ZL, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 2015;72(1):137–51 (1).
[4] Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn 2012;30(2):137–49.
[5] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337(3):635–45.
[6] Liu ZR, Huang YQ. Advantages of proteins being disordered. Protein Sci 2014;23 (5):539–50.
[7] Uversky VN. Intrinsic disorder-based protein interactions and their modulators. Curr Pharm Des 2013;19(23):4191–213.
[8] Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18(5):343–84 (5).
[9] Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. BMC Genom 2009;10 (Suppl 1):S12.
[10] Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 2008;37:215–46.
[11] Kulkarni P, Uversky VN. Intrinsically disordered proteins in chronic diseases. Biomolecules 2019;9(4).
[12] Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochem Soc Trans 2016;44(5):1185–200.
[13] Su BG, Henley MJ. Drugging fuzzy complexes in transcription. Front Mol Biosci 2021;8:795743.
[14] Biesaga M, Frigole-Vivas M, Salvatella X. Intrinsically disordered proteins and biomolecular condensates as drug targets. Curr Opin Chem Biol 2021;62:90–100.
[15] Hosoya Y, Ohkanda J. Intrinsically disordered proteins as regulators of transient biological processes and as untapped drug targets. Molecules 2021;26(8).
[16] Ghadermarzi S, Li S, Li M, Kurgan L. Sequence-derived markers of drug targets and potentially druggable human proteins. Front Genet 2019;10:1075.
[17] Hu G, Wu Z, Wang K, Uversky VN, Kurgan L. Untapped potential of disordered proteins in current druggable human proteome. Curr Drug Targets 2016;17(10): 1198–205.
[18] Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A, Consortium D, Tosatto SCE, Piovesan D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res 2023.
[19] Piovesan D, Del Conte A, Clementel D, Monzon AM, Bevilacqua M, Aspromonte MC, et al. MobiDB: 10 years of intrinsically disordered proteins. Nucleic Acids Res 2023;51(D1):D438–44.
[20] Bateman A, Martin Maria-Jesus, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51 (D1):D523–31.
[21] Atkins JD, Boateng SY, Sorensen T, McGuffin LJ. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. Int J Mol Sci 2015;16(8):19040–54.
[22] He B, Wang K, Liu Y, Xue B, Uversky VN, Keith Dunker A. Predicting intrinsic disorder in proteins: an overview. Cell Res 2009;19(8):929–49.
[23] Kurgan L, Li M, Li Y. The methods and tools for intrinsic disorder prediction and their application to systems medicine. In: Wolkenhauer O, editor. Systems Medicine. Academic Press: Oxford; 2021. p. 159–69.
[24] Meng F, Uversky V, Kurgan L. Computational prediction of intrinsic disorder in proteins. Curr Protoc Protein Sci 2017;88. p. 2 16 1-2 16 14.
[25] Zhao B, Kurgan L. Deep learning in prediction of intrinsic disorder in proteins. Comput Struct Biotechnol J 2022;20:1286–94.
[26] Zhao B, Kurgan L. Surveying over 100 predictors of intrinsic disorder in proteins. Expert Rev Proteom 2021;18(12):1019–29.
[27] Kurgan L, Hu G, Wang K, Ghadermarzi S, Zhao B, Malhis N, et al. Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins. Nat Protoc 2023;18(11):3157–72.
[28] Kurgan L. Resources for computational prediction of intrinsic disorder in proteins. Methods 2022;204:132–41.
[29] Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief Bioinform 2019;20(1):330–46.
[30] Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003; 53(Suppl 6):561–5.
[31] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins-Struct Funct Bioinforma 2014;82:1–6.
[32] Necci M, Piovesan D, Predictors CAID, Curators DisProt, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods 2021;18(5). 472-+.
[33] Del Conte A, Mehdiabadi M, Bouhraoua A, Monzon AM, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. Proteins-Struct Funct Bioinforma 2023.
[34] Zhao B, Kurgan L. Deep learning in prediction of intrinsic disorder in proteins. Comput Struct Biotechnol J 2022;20:1286–94.
[35] Katuwawala A, Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions. Brief Bioinform 2020;21(5):1509–22.
[36] Necci M, Piovesan D, Dosztányi Z, Tompa P, Tosatto SCE. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics 2018;34(3):445–52.
[37] Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SCE. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 2015;31(2):201–8.
[38] Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012;13(1):6–18.
[39] Zhao B, Ghadermarzi S, Kurgan L. Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. Comput Struct Biotechnol J 2023;21:3248–58.
[40] Hu G, Wang K, Song J, Uversky VN, Kurgan L. Taxonomic landscape of the dark proteomes: whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity. Proteomics 2018:e1800243.
[41] Wang C, Uversky VN, Kurgan L. Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from eukaryota, bacteria and archaea. Proteomics 2016;16(10):1486–98.
[42] Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biol 2011;12(12):R120.
[43] DeForte S, Uversky VN. Not an exception to the rule: the functional significance of intrinsically disordered protein regions in enzymes. Mol Biosyst 2017;13(3):463–9.
[44] Necci M, Piovesan D, Tosatto SC. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. Protein Sci 2016;25(12): 2164–74.
[45] Xue B, Williams RW, Oldfield CJ, Keith Dunker A, Uversky VN. Archaic chaos: intrinsically disordered proteins in Archaea. BMC Syst Biol 2010;4(Suppl 1):S1.
[46] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberge O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873): 583–9.
[47] Hanson J, Paliwal K, Zhou Y. Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures. J Chem Inf Model 2018;58(11):2369–76.
[48] Wang S, Ma JZ, Xu JB. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 2016;32(17): 672–9.
[49] Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, Kurgan L, et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun 2021;12(1):4438.
[50] Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012;28(4):503–9.
[51] Mirabello C, Wallner B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. Plos One 2019;14(8).
[52] Orlando G, Raimondi D, Codicè F, Tabaro F, Vranken W. Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. J Mol Biol 2022;434(12):167579.
[53] Hanson J, Paliwal KK, Litfin T, Zhou Y. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. Genom Proteom Bioinforma 2019; 17(6):645–56.
[54] Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. Protein Sci 2022;31(11).
[55] Wilson CJ, Choy WY, Karttunen M. AlphaFold2: a role for disordered protein/ region prediction? Int J Mol Sci 2022;23(9).
[56] Zhao B, Ghadermarzi S, Kurgan L. Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. Comput Struct Biotechnol J 2023;21:3248–58.
[57] Kabsch W, Sander C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22 (12):2577–637.
[58] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50 (D1):D439–44.
[59] Perrigo, B. Google's AI Lab, DeepMind, Offers 'Gift to Humanity' with Protein Structure Solution. 2022 [cited 2024 April 5]; Available from: https://time.com/6201423/deepmind-alphafold-proteins/.
[60] Piovesan D, Del Conte A, Clementel D, Miguel Monzon A, Bevilacqua M, Aspromonte MC, et al. MobiDB: 10 years of intrinsically disordered proteins. Nucleic Acids Res 2023;51(D1):D438–44.

[61] Di Domenico T, Walsh I, Martin AJM, Tosatto SCE. MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 2012;28(15):2080–1.
[62] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res 2023;51(D1):D488–508.
[63] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–2.
[64] Yan J, Mizianty MJ, Filipow PL, Uversky VN, Kurgan L. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochim Biophys Acta 2013;1834(8):1671–80.
[65] Skupien-Rabian B, Jankowska U, Swiderska B, Lukasiewicz S, Ryszawy D, Dziedzicka-Wasylewska M, Kedracka-Krok S. Proteomic and bioinformatic analysis of a nuclear intrinsically disordered proteome. J Proteom 2016;130:76–84.
[66] Zhang H, Zhang T, Chen K, Kedarisetti KD, Mizianty MJ, Bao Q, et al. Critical assessment of high-throughput standalone methods for secondary structure prediction. Brief Bioinform 2011;12(6):672–88.
[67] Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34(4):508–19.
[68] Guo ZY, Hou J, Cheng JL. DNSS2: Improved ab initio protein secondary structure prediction using advanced deep learning architectures. Proteins-Struct Funct Bioinforma 2021;89(2):207–17.
[69] Tamposis IA, Sarantopoulou D, Theodoropoulou MC, Stasi EA, Kontou PI, Tsirigos KD, Bagos PG. Hidden neural networks for transmembrane protein topology prediction. Comput Struct Biotechnol J 2021;19:6090–7.
[70] Zhang Y, Sagui C. Secondary structure assignment for conformationally irregular peptides: comparison between DSSP, STRIDE and KAKSI. J Mol Graph Model 2015;55:72–84.
[71] Liu T, Wang Z. SOV_refine: A further refined definition of segment overlap score and its significance for protein structure similarity. Source Code Biol Med 2018;13.
[72] Kurgan L, Disfani FM. Structural protein descriptors in 1-dimension and their sequence-based predictions. Curr Protein Pept Sci 2011;12(6):470–89.
[73] Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. Intrinsically Disord Proteins 2016;4(1):e1259708.
[74] Zhao B, Kurgan L. Compositional bias of intrinsically disordered proteins and regions and their predictions. Biomolecules 2022;12(7).
[75] Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Keith Dunker A. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett 2008;15(9):956–63.
[76] Williams RM, Obradovic Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunjker AK. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pac Symp Biocomput 2001:89–100.