Resources for computational prediction of intrinsic disorder in proteins

Lukasz Kurgan[1*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, United States

*Corresponding author: Lukasz Kurgan (lkurgan@vcu.edu)

Abstract

With over 40 years of research, intrinsic disorder prediction field has developed over 100 computational predictors. This review offers a holistic perspective of this field by highlighting accurate and popular disorder predictors and introducing a wide range of practical resources that support collection, interpretation and application of disorder predictions. These resources include meta webservers that expedite collection of multiple disorder predictions, large databases of pre-computed disorder predictions that ease collection of predictions particularly for large datasets of proteins, and modern quality assessment tools. The latter methods facilitate identification of accurate predictions in a specific protein sequence, reducing uncertainty associated to the use of the putative disorder. Altogether, we review eleven predictors, four meta webservers, three databases and two quality assessment tools, all of which are conveniently available online. We also offer a perspective on future developments of the disorder prediction and the quality assessment tools. The availability of this comprehensive toolbox of useful resources should stimulate further growth in the application of the disorder predictions across many areas including rational drug design, systems medicine, structural bioinformatics and structural genomics.

Keywords: intrinsic disorder; intrinsically disordered proteins; machine learning; webserver; prediction; protein function.

1 Introduction

Intrinsic disorder is manifested by presence of regions in protein sequences that are absent of a well-defined equilibrium structure under physiological conditions [1-4]. Intrinsically disordered proteins (IDPs) include one or more intrinsically disordered regions (IDRs) interspersed with structured regions, and in some cases, they are composed of one sequence-wide IDR. Several bioinformatics studies suggest that IDPs are relatively common in all kingdoms of life and viruses [5-9]. Some studies predict that about 1/3 of eukaryotic proteins have long IDRs, i.e., sequence regions composed of 30 or more consecutive disordered residues [5, 6, 10]. IDPs are instrumental for numerous cellular functions including molecular assembly and recognition, cell cycle regulation, signaling, transcription and translation, to name a few [11-23]. They are found across several membrane-bound cellular compartments [24, 25] and are important for the formation of membraneless compartments via the liquid-liquid phase separation [26-28]. They are also central to the formation of dark proteomes, defined as proteins that are not amenable to commonly used methods of experimental structure determination, such as X-ray crystallography [29-32].

Several databases, such as DisProt [33-37], PDB [38], IDEAL [39], DIBS [40], FuzDB [41] and MFIB [42], provide access to experimental data on IDPs. However, these resources cover a small sliver of the protein sequence space. More specifically, version 9.01 of DisProt includes about 2,400 IDPs [37] and a recent study identified approximately 25,000 IDPs in PDB [43], compared to over 220 million proteins in the RefSeq database [44]. This large and growing annotation gap motivates development of accurate computational methods that predict disordered residues and regions in a given protein sequence. The underlying ability to produce accurate predictions stems from the observation that disorder is an inherent/intrinsic property of the input amino acid sequences [4]. Moreover, the development and applications of the bioinformatics methods, including the disorder predictors, was shown to stimulate rapid acceleration in the research on IDPs and IDRs [45].

A recent study finds that over 100 disorder predictors were developed to date [46]. Numerous surveys summarize these predictors [43, 46-56]. These articles offer a historical perspective, describe and classify selected collections of disorder predictors and, in some cases, compare their predictive quality. The comparative studies offer invaluable advice on how to select the most accurate methods [43, 48, 53, 54, 57-63]. However, these surveys and comparative studies miss the opportunity to introduce other important resources that facilitate use and applications of disorder predictors. Nowadays, users have the options to conveniently collect pre-computed disorder predictions, generate multiple disorder predictions with one request and utilize disorder quality assessment tools. We provide a holistic perspective of the disorder prediction field. We summarize a selected collection of popular and/or accurate disorder predictors, identify several webservers that offer multiple disorder predictions, introduce and contrast large databases of putative disorder, and discuss how to obtains useful insights into correctness of the disorder predictions using modern disorder quality assessment tools.

2 Computational predictors of intrinsic disorder

Prior surveys identify six disorder predictors that were released between 1979, when the first method was published [64], and 2002 [46, 52, 55]. The development efforts have accelerated after 2002, which coincides with the inclusion of the disorder prediction assessment into the Critical Assessment of protein Structure Prediction (CASP) experiment in 2002 [62]. Since then, on average five new methods are developed annually [46], including at least nine predictors that were published since 2020: DisoMine [65], ODiNPred [66], IDP-Seq2Seq [67], flDPnn [68], flDPlr [68], IUPred3 [69], RFPR-IDP [70], MetaPredict [71], and DeepCLD [72]. Recent survey finds that majority of recent disorder predictors rely on deep neural network models [46], which partly stems from the success of deep learners in recent disorder prediction assessments [48, 56, 57]. Here, we focus on providing practical advice for the end users by identifying a selection of popular and/or most accurate disorder predictors.

About a dozen large-scale comparative assessments of the intrinsic disorder predictors were carried out so far [43, 48, 53, 54, 57-63]. They include several community assessments where disorder predictors are tested on blind datasets by independent assessors. These datasets are withheld from the authors of the predictors before the assessment (i.e., authors are blind to the content of these datasets) and the assessors do not participate in the competitions, arguably making these evaluations more objective when compared to the comparative studies done by the authors of predictors. The disorder prediction was included in the biannual CASP experiment between CASP4 (2002) and CASP10 (2012) [58-63], and it was discontinued after CASP10. Another community assessment, Critical Assessment of Intrinsic Protein Disorder (CAID), was published in 2021 [57]. We focus on the two most recent community assessments, CASP10 and CAID to identify a pool of well-performing disorder predictors. We include the two assessments since they utilize benchmark data from two complementary sources, PDB depositions in CASP10 and DisProt records in CAID.

To quote conclusions from the CASP10 assessment: "*Four prediction groups – Prdos-CNF, DISOPRED3, biomine_dr_mixed, and biomine_dr_pdb_c – perform better than the others according to the majority of the evaluation measures.*" [61]. The corresponding top-three disorder predictors are PrDOS [73], DISOPRED3 [74], and MFDp [75-77], given that biomine_dr_mixed, and biomine_dr_pdb_c predictions cover two versions of MFDp where the better performing version is implemented by the MFDp tool. Similarly, the results from CAID are perhaps best summarized by the quote from the accompanying commentary [78]: "*SPOT-Disorder2 and flDPnn, followed by RawMSA and AUCpreD, are consistently good. However, flDPnn is at least an order of magnitude faster than its competitors, and it succeeded on all sequences, whereas SPOT-Disorder2 skipped 5% of sequences as a* *result of a length limitation. This might make fIDPnn the overall winner of CAID, but since fIDPnn and RawMSA are not yet published, SPOT-Disorder2 and AUCpreD are the best options available to researchers today.*" The flDPnn and RawMSA methods were published in the meantime and thus the four methods, which include AUCpreD [79], rawMSA [80], SPOT-Disorder2 [81], and flDPnn [68], are available to the end users.

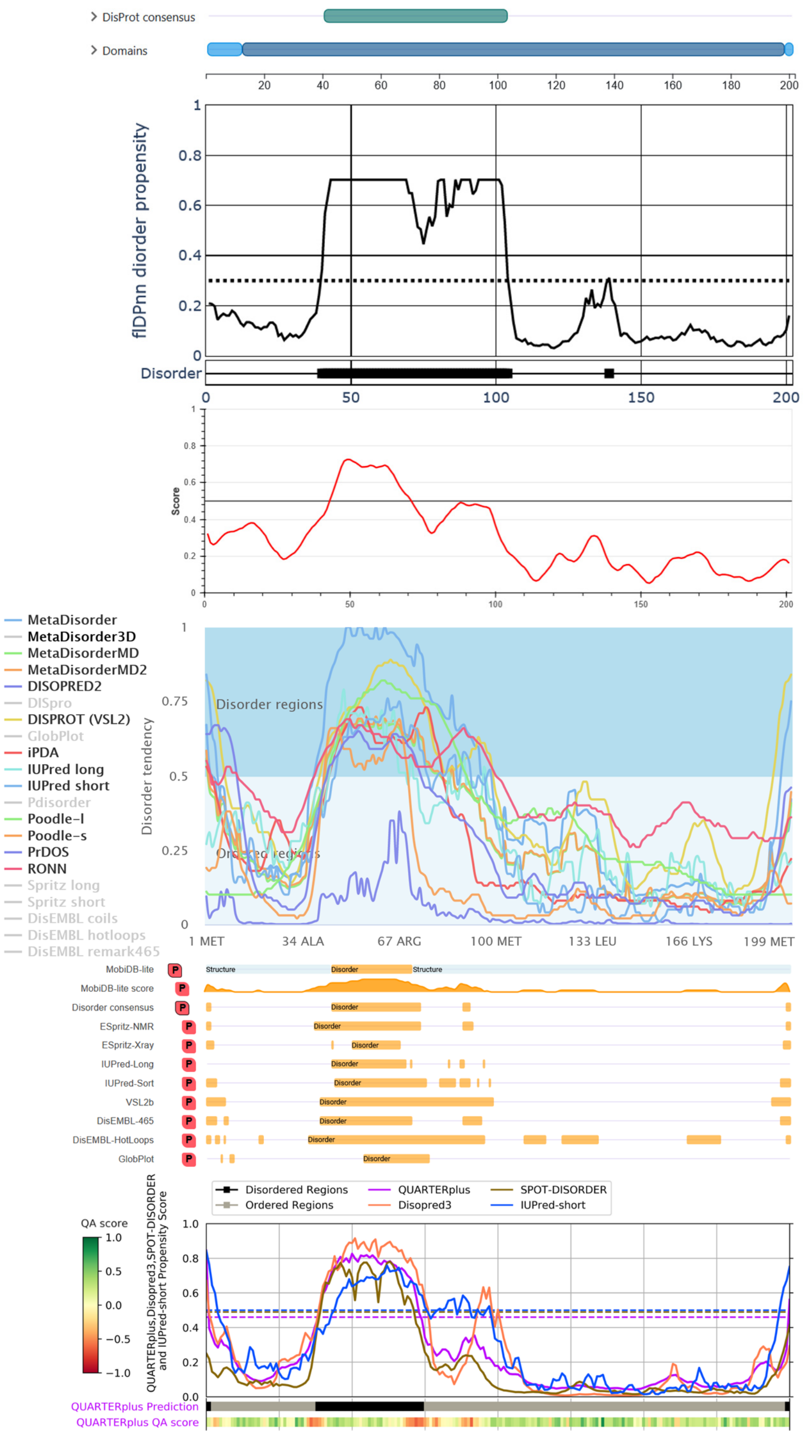
We note that the CAID benchmark was also recently used to investigate potential impact of results produced by AlphaFold2 [82], which has disrupted the protein structure prediction area, on the disorder prediction field. Two recently released manuscripts use the CAID dataset to assess whether the AlphaFold2's predictions can rival current disorder predictors [83, 84]. They conclude that while AlphaFold2's results can be used to identify disordered regions, they are outperformed by the modern disorder predictors that include the best methods from the CAID experiment. This suggests that disorder and structure prediction areas require different solutions and they effectively complement each other.

We supplement the above list of the seven accurate disorder predictors, three from CASP10 and four from CAID, with a few popular methods. We quantify popularity with the citations collected using Google Scholar in January 2022. We normalize the total number of citations by the time since the corresponding articles were published, measured in months. The four highly cited methods include (in the order of their citation rates): IUPred [69, 85-87], DISOPRED [74, 88, 89], PONDR VSL2 [90, 91], and DisEMBL [92] (Table 1). We use the recent CAID results to provide context of the differences in the predictive performance between the accurate and the popular predictors. Using the results for the DisProt dataset from CAID [57], the F1 metric values of the accurate predictors that were included in this experiment are 0.48 (flDPnn), 0.47 (SPOT-Disorder2), 0.45 (RawMSA), 0.43 (AUCpreD) and 0.39 (DISOPRED3). To compare, the corresponding F1 scores for the popular methods are 0.42 (IUPred), 0.41 (PONDR VSL2) and 0.36 (DisEMBL). While these results show that certain methods produce accurate predictions of disorder, they are not necessarily equally good when applied in the context of related predictions. A recent report analyzes predictive performance of ten disorder predictors applied to identify protein-binding and nucleic acid-binding regions [48]. This study finds that the predictions from SPOT-Disorder are the best for the identification of nucleic acid-binding while the ESpritz method [93] produces disorder that works well for the prediction of protein-binding.

We summarize the 11 most accurate and/or popular disorder predictors in Table 1. We list the citation data, types of predictive models that are used by these methods, and websites where these methods can be accesses. The most cited methods are published in mid 2000s and they rely on traditional machine learning models, such as shallow feed-forward neural networks (DISOPRED and DisEMBL) or support vector machines (PONDR VSL2 and MFDp), and scoring functions that approximate physical principles governing protein folding (IUPred). In contrast, the recent CAID-winning predictors utilize more sophisticated deep learning models. We note that these deep learners utilize several different neural network architectures including the deep feed-forward topology (flDPnn), recurrent topology (SPOT-Disorder2), and a hybrid architecture that combines convolutional and recurrent elements (RawMSA and AUCpred). Importantly, the 11 selected disorder predictors are freely available to the end users via the websites shown in Table 1. Most of these methods, including DISOPRED, DisEMBL, IUPred, SPOT-Disorder2 and flDPnn are provided in two complementary ways: as a standalone code that can be installed and executed on the user's hardware and a webserver that conveniently runs on the server-side without the need for installation. Four predictors, PONDR VSL2, PrDOS, MFDp and AUCpred are available as webservers, while rawMSA is offered solely as the standalone code. Altogether, we conclude that users have access to a variety of readily available, popular and accurate options.

Table 1. Popular and accurate predictors of intrinsic disorder. The methods are sorted in the chronological order of their publication. The citations were collected from Google Scholar as of January 2022. Total citations number is calculated as the sum over all listed publications. The per month citations number is the total number of citations divided by the number of months since the first publication. Predictive models used include SVM (support vector machine), SFFNN (shallow feed-forward neural network), and DNN (deep neural network).

Predictor	First published	Ref	Reason for inclusion	Total citations (per month)	Predictive model used	URL
DISOPRED	Oct. 2003	[74, 88, 89]	High citation rate	1,593 (7.3)	SFFNN	http://bioinf.cs.ucl.ac.uk/psipred/
DisEMBL	Nov. 2003	[92]	High citation rate	1,377 (6.3)	SFFNN	http://dis.embl.de/
IUPred	April 2005	[69, 85-87]	High citation rate	3,633 (18.1)	Scoring function	https://iupred1.elte.hu/
PONDR VSL2	Sept. 2005	[90, 91]	High citation rate	1,357 (6.9)	SVM	http://www.pondr.com/
PrDOS	July 2007	[73]	High CASP10 rank	681 (3.9)	SVM	https://prdos.hgc.jp/cgi-bin/top.cgi
MFDp	Sept. 2010	[75-77]	High CASP10 rank	293 (2.2)	SVM	http://biomine.cs.vcu.edu/servers/MFDp/
DISOPRED3	March 2015	[74]	High CASP10 rank	573 (7.0)	Ensemble of SVM, SFFNN and nearest neighbor	http://bioinf.cs.ucl.ac.uk/psipred/
AUCpred	Sept. 2016	[79]	High CAID rank	66 (1.0)	DNN (hybrid convolutional and conditional random field)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/
rawMSA	Aug. 2019	[80]	High CAID rank	23 (0.8)	DNN (hybrid convolutional and recurrent)	https://bitbucket.org/clami66/rawmsa/src/master/
SPOT-Disorder2	Dec. 2019	[81]	High CAID rank	43 (1.7)	DNN (recurrent)	https://sparks-lab.org/server/spot-disorder2/
flDPnn	July 2021	[68]	High CAID rank	8 (1.3)	DNN (feed-forward)	http://biomine.cs.vcu.edu/servers/flDPnn/



A. Experimental disorder annotation from DisProt.

B. Disorder predicted by flDPnn webserver.

C. Disorder predicted by IUPred3 webserver.

D. Disorder predicted by MetaDisorder meta webserver.

E. Putative disorder collected from the MobiDB database.

F. Putative disorder and accompanying quality assessment (QA) scores generated using the QUARTERplus webserver.

Figure 1. Experimental and predicted disorder annotations for the 50S ribosomal protein L4 (UniProtKB id: P60723). Panel A shows experimental annotations of an IDR (positions 41 to 103) collected from DisProt (DisProt id: DP00600) (https://www.disprot.org/). Panel B shows putative disorder produced by flDPnn (http://biomine.cs.vcu.edu/servers/flDPnn/), where black plot gives the putative disorder propensity and horizonal black bar corresponds to putative IDR derived from propensity values using the cutoff denoted with the dotted line. Panel C is the putative disorder generated by IUPred3(https://iupred3.elte.hu/), where red plot shows the putative disorder propensity and horizonal black line is the cutoff used to identify putative IDRs. Panel D shows disorder predictions generated by multiple color-coded methods that are included in the MetaDisorder webserver (http://iimcb.genesilico.pl/metadisorder/), where the darker blue background identifies the threshold used to derive IDRs. Panel E provides predictions of disorder that are available in MobiDB (https://mobidb.bio.unipd.it/), where the horizonal orange bars denote predicted IDRs. Panel F gives putative disorder generated by a consensus of SPOT-DISORDER-Single [94], DISOPRED3 [74] and IUPred-short [86] (black and gray horizontal bar line) that is accompanied by the quality assessment (QA) scores produced by QUARTERplus (http://biomine.cs.vcu.edu/servers/QUARTERplus/). The QA scores (color-coded horizonal bar) quantify predictive quality of the consensus disorder prediction, i.e., amino acids shown in green and yellow colors are more likely to be accurately predicted compared to the residues colored in orange or red.

We illustrate results that these methods generate based on an example that predicts disorder for the 50S ribosomal protein L4 (UniProtKB id: P60723). This 201-residues long protein facilitates early stages of the ribosome assembly and acts as a transcriptional repressor [95, 96]. Experimental annotations of disorder, which we collect from DisProt (DisProt id: DP00600) [37], reveal that the L4 protein harbors a long IDR (positions 41 to 103) that is engaged during the ribosome assembly process [96, 97]; see Figure 1A. Using data from Table 1, we show predictions generated by the most recent method, flDPnn (Figure 1B), and the most cited predictor, IUPred (Figure 1C). The disorder predictions are composed of two parts: numerical propensity scores that quantify likelihood for disorder for each amino acid in a given protein sequences, and binary scores that annotate each residue as disordered vs. ordered. The binary predictions typically form regions (IDRs) and are computed from the putative propensities such that amino acids with propensities greater than a threshold are designated as disordered. The thresholds used by flDPnn and IUPred are shown using black horizontal dotted line (Figure 1B) and black horizontal solid line (Figure 1C), respectively. We find that flDPnn correctly identifies the presence and location of the IDR in this protein (Figure 1B). Similarly, IUPred also generates relatively high propensities for residues that coincide with the location of the experimental IDR (Figure 1C).

Table 2. Meta webservers that provide at least four disorder predictions generated by different methods. The webservers are sorted in the chronological order of their publication. We use bold font to highlight the popular and/or accurate methods from Table 1.

Webserver	First published	Ref	Disorder predictors included	URL
MetaPrDOS	June 2008	[98]	**PrDOS** [73], **DISOPRED2** [5], **DisEMBL** [92], **PONDR VSL2** [90], DISpro [99], **IUpred** [85, 86], POODLE-S [100], DISOclust [101], and metaPrDOS [98]	https://prdos.hgc.jp/meta/about.html
MFDp	Sept. 2010	[75-77]	DISOclust [101], **DISOPRED** [88, 89], **IUPred** [85, 86], and **MFDp** [75]	http://biomine.cs.vcu.edu/servers/MFDp/
MetaDisorder	May 2012	[102]	**DisEMBL** [92], **DISOPRED2** [5], DISpro [99, 103], GlobPlot [104], iPDA [105], **IUPred** [85, 86], Pdisorder, POODLE-S [100], POODLE-L [106], **PrDOS** [73], Spritz [107], and RONN [108]	http://iimcb.genesilico.pl/metadisorder/
DisMeta	Oct. 2013	[109]	**DisEMBL** [92], **DISOPRED2** [5], DISpro [99], FoldIndex [110], GlobPlot [104], **IUPred** [85, 86], RONN [108], and **PONDR VSL2** [90]	http://montelionelab.chem.rpi.edu/dismeta/

3 Meta webservers for disorder prediction

We examine recent surveys [43, 46, 48, 50-52, 54, 56] and manually screen results of a PubMed search using the "prediction[Title] AND intrinsic AND disorder" query to identify online resources that provide access to multiple disorder predictions. We require that these resources generate and output results of at least four different intrinsic disorder predictors. The key benefit of having access to multiple results is that they can be assessed for convergence in order to boost confidence in the resulting disorder prediction. The latter is motivated by a few studies that empirically demonstrate that the use of a consensus-based disorder prediction typically leads to an improved predictive performance [43, 111-114]. We found four meta webservers, which we summarize in Table 2. They include MetaDisorder [102], MetaPrDOS [98], DisMeta [109] and MFDp [75-77] that generate 12, 9, 8 and 4 disorder predictions, respectively. Availability via the webserver is especially valuable for users who perform predictions in an *ad hoc* manner and less computer savvy users. Users simply input their protein sequence(s) via a web browser using the URL listed in Table 2, and collect the resulting predictions that are delivered via the browser window and/or via email. The prediction process that covers running multiple methods is automated and executed on the server side. This means that users do not need to use their own hardware or install software to collect the results, which makes these resources very convenient. Table 2 reveals that these meta webservers generate predictions for the four popular methods (DISOPRED, DisEMBL, IUPred and PONDR VSL2) and two of the top-performers from CASP10 (PrDOS and MFDp). However, since these webservers were released and published several years ago, they do not include the newer accurate predictors, such as DISOPRED3, AUCpred, rawMSA, SPOT-Disorder2 and flDPnn.

Figure 1D shows predictions for the L4 protein produced by MetaDisorder, which is the resource that covers the largest number of disorder predictors. Propensity scores output by different predictors are color-coded and they reveal that majority of the methods converge on predicting residues in the region spanning between positions 40 and 80 as disordered. This prediction overlaps with the location of the experimentally-found IDR (Figure 1A). While some methods predict disorder at the sequence termini, it is clear that these methods are in minority, which suggests that this prediction should be disregarded (Figure 1D). This illustrates how the consensus analysis can be used to guide the process of identifying IDRs.

4 Databases of intrinsic disorder predictions

Collection of the disorder predictions could be difficult and time consuming, especially when assembling results of multiple methods that are not available via one of the meta webservers. In that case, users must navigate multiple websites and/or install multiple software, deliver inputs (i.e., sequences, identifiers and/or emails) in multiple formats, and parse and standardize the different formats of outputs that disorder predictors use. Another substantial drawback is that the runtime required to make predictions could be substantial, as much as several minutes for one protein [57]. This is particularly challenging when collecting predictions for large protein datasets, such as big protein families or proteomes. We note that studies that perform proteome-wide analysis of disorder are done regularly [5-7, 24, 29, 115-120].

A suitable alternative to the direct use the disorder predictors is to collect pre-computed disorder predictions from one of the three available databases: Database of Disorder Protein Predictions (D^2P^2) [121], MobiDB [122-125], and DescribePROT [126]. These resources provide access to disorder prediction for large collections of proteins ranging from 1.37 million in DescribePROT, 10.43 million proteins in D^2P^2, to 219.74 million proteins in MobiDB (Table 3). The key feature of these databases is the instantaneous retrieval of the pre-computed predictions generated by multiple methods. Users do not have to wait for the computation of predictions and do not need to assemble the various predictions together. Moreover, this reduces wasteful duplication in the use of the predictors that are often tasked to make many predictions for the same protein when the same sequence is submitted by different users. The three databases facilitate collection of predictions for individual proteins, which are provided in a parsable text format and in an interactive graphical format (Figure 1E). They also provide options to conveniently download predictions for whole proteomes. For example, DescribePROT offers access to the raw predictions (propensities and binary values) and protein-level summaries that include disorder content at the whole proteome level (http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/download.html). MobiDB and D^2P^2 databases deliver predictions from 8 and 6 disorder predictors (Table 3), respectively, and also produce a consensus result. MobiDB computes the consensus using the MobiDB-lite algorithm [113] and D^2P^2 uses the 75% consensus approach, i.e., an amino acid is predicted as disorder if at least 75% of methods predicts it as disordered. Figure 1E shows disorder predictions for the L4 protein that we collect from MobiDB, which covers the largest number of disorder predictors. The first two lines give the binary predictions and the corresponding propensities generated by MobiDB-lite. The subsequent lines provide binary predictions from the individual disorder predictors. We note a good agreement between these predictions, in particular the consensus-based result, and the corresponding experimental data from Figure 1A.

Table 3. Databases of intrinsic disorder predictions. The databases are sorted in the chronological order of their publication. We use bold font to highlight the popular and/or accurate methods from Table 1.

Database	First published	Ref	Size [millions proteins]	Disorder predictors included	Other predictions included	URL
MobiDB version 4.1	Aug. 2012	[124]	219.74	**DisEMBL** [92], DynaMine [127], ESpritz [93], GlobPlot [104], **IUPred2A** [87], JRONN [108], MobiDB-lite [113], **PONDR VSL2** [90, 91]	Disordered protein-binding by ANCHOR [128] Secondary structure by FeSS [129] Low complexity regions by SEG [130] Domains by Gene3D [131]	https://mobidb.bio.unipd.it/
D^{2}P^2 version 1.0	Jan. 2013	[121]	10.43	PONDR VL-XT [132], **PONDR VSL2** [90, 91], **PrDOS** [98], PV2 [133], ESpritz [93], **IUPred** [86]	Disordered protein-binding residues by ANCHOR [128] Domains by SUPERFAMILY [134]	https://d2p2.pro/
DescribePROT version 1.4	Jan. 2021	[126]	1.37	**PONDR VSL2** [90, 91]	Solvent accessibility by ASAquick [135] Secondary structure by PSIPRED [136-138] Disordered DNA- and RNA-binding by DisoRDPbind [139-141] Disordered protein-binding by DisoRDPbind [139-141] and MoRFChibi [142] Structured protein-binding by SCRIBER [143, 144] Structured RNA-binding and DNA-binding by DRNApred [145] Disordered DNA-binding by Disordered linkers by DFLpred [146] Signal peptides by SignalP [147, 148]	http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/

We also point to the key advantages and drawbacks of these resources. MobiDB covers by far the largest number of proteins. It is also cross-linked and includes experimental data from ten external sources: CoDNaS [149], DIBS [40], DisProt [33], ELM [150], FuzDB [41], IDEAL [39], MFIB [42], PDBe [151], PhasePro [152], and UniProt [153]. On the other hand, MobiDB is almost exclusively focused on the intrinsic disorder. D^2P^2 offers arguably the most comprehensive annotations of protein domains and posttranslational modification sites that are collected using SUPERFAMILY [134] and PhosphoSitePlus [154], respectively. However, it is also heavily focused on the intrinsic disorder and is no longer maintained. It was last updated in 2013. DescribePROT provides access to the predictions for a diverse set of structural and functional characteristics of proteins, including intrinsic disorder. The other characteristics consist of alignment profiles and putative solvent accessibility, disordered linkers, secondary structure, signal peptides, and protein-binding, DNA-binding and RNA-binding residues. Predictors used to derive these annotations are listed in Table 3. Altogether, release 1.4 of DescribePROT provides access to over 7.8 billion amino acid-level predictions. However, DescribePROT covers the smallest number of proteins and only one disorder predictor.

Finally, while undoubtedly these are very helpful resources, they do not offer a holistic solution. In particular, users who would like to collect result outside of the protein sets covered in these databases, e.g., for a novel protein sequence, have to rely on the disorder predictors and meta webservers

5 Methods for quality assessment of disorder predictions

The comparative assessments of disorder predictors, including CASP10 and CAID, report results on datasets composed of dozens or hundreds of proteins. This quantifies an overall, dataset-wide predictive performance of these methods. However, these studies may not offer an accurate guidance when predicting specific proteins. A recent study shows that predictive performance of disorder predictors varies widely between proteins, where some sequences are predicted very accurately while other predictions are hardly better than random [56]. Other works find that the predictive quality varies between amino acids in the same sequences and that the outputs generated by the predictors do not allow to accurately identify these differences [155, 156]. To this end, one of the more recent advances in the disorder prediction field is the development and release of methods that provide quality assessment (QA) scores [157]: QUARTER [156, 158] and its upgraded version QUARTERplus [159]. The QA scores quantify confidence in the disorder predictions at an amino acid level [156]. In other words, these scores facilitate identification of amino acids in a given protein chain for which the disorder predictions are more likely to be accurate.

QUARTERplus webserver (http://biomine.cs.vcu.edu/servers/QUARTERplus/) produces disorder predictions and the associated QA scores for several disorder predictors including DISOPRED3 [74], IUPred [85, 86], PONDR VSL2 [90, 91], DisEMBL [92], GlobPlot [104], and SPOT-Disorder-Single [94]. Moreover, it uses a deep convolutional neural network to make accurate, consensus-based disorder predictions using outputs from SPOT-Disorder-Single, IUPred and DISOPRED3 that are accompanied by the QA scores. These QA scores allow users to conveniently pinpoint which disorder predictions are more trustworthy. We illustrate this in Figure 1F where we present the consensus disorder prediction produced by QUARTERplus (pink line and black and gray horizontal bar) for the L4 protein. This prediction is coupled with the color-coded QA scores where green and yellow denote residues that are more likely to be accurately predicted compared to the predictions colored in orange or red. The usefulness of the QA scores becomes apparent when comparing the disorder predictions from QUARTERplus (Figure 1F) and the corresponding experimental annotations (Figure 1A). While QUARTERplus predicts a short IDR at the N-terminus, the associated QA scores are orange, which suggests that this is a false prediction, in agreement with the native annotations. The long experimental IDR located between positions 41 and 103 is underpredicted by QUARTERplus around positions 80 to 103. However, the corresponding QA scores are colored red and yellow, which correctly suggests low quality of these disorder predictions. At the same time, the QA scores for the positions between 103 and the C-terminus are green and yellow, which indicate that the related disorder predictions are likely accurate. This agrees with the experimental annotations for this part of the sequence (Figure 1A). This example demonstrates how the QA scores can be used to identify correct vs. incorrect disorder predictions for a given protein sequence. Altogether, we conclude that QUARTERplus a is a convenient tool which provides QA scores that empower users to identify more trustworthy predictions for several popular disorder predictors.

6 Summary and outlook

Intrinsic disorder prediction field produced over 100 predictors [46]. We highlight 11 disorder predictors selected based on the results of two most recent community assessments, CASP10 [61] and CAID [57], and citation data. This collection represents arguably the most accurate and most popular methods at this time. More importantly, we introduce and discuss a wide range of other practical resources including meta webservers, databases and disorder QA tools. These resources facilitate convenient collection, interpretation and application of disorder predictions.

The meta webservers, which include MetaDisorder [102], MetaPrDOS [98], DisMeta [109] and MFDp [75-77], expedite collection of multiple disorder predictions, which can be effectively used to perform a consensus-based analysis. The users have an option to obtain pre-computed disorder predictions from three large databases, such as MobiDB [125], D^2P^2 [121], and DescribePROT [126]. This is particularly valuable when targeting analysis of disorder for large datasets of proteins and when analyzing disorder in the context of other structural and functional features, such as protein domains (available in D^2P^2 and MobiDB), posttranslational modification sites (D^2P^2), secondary structure (MobiDB and DescribePROT), protein-protein interactions (MobiDB and DescribePROT), and protein-nucleic acid interactions, linkers and solvent accessibility (DescribePROT). Moreover, we discuss QUARTERplus [159], a modern deep learning-based tool that produces QA scores for several popular disorder predictors. The QA scores facilitate identification of accurate predictions in a specific protein sequence, reducing uncertainty associated with the use of putative disorder annotations. Availability of this comprehensive toolbox of practical resources will fuel further growth of the computational intrinsic disorder field, ensuring that the disorder predictions will continue to make impact across many application areas, such as drug design [160-163], systems medicine [164], structural bioinformatics [143, 165-169], and structural genomics [29, 32, 92]. In the context of the rational drug design, the disorder predictions are used to identify currently underrepresented disordered proteins as drug targets [160-162, 170] and to facilitate the development of new computational tools for modelling protein-drug interactions [171-174]. As another example, the putative annotations of intrinsic disorder facilitate target selection in structural genomics by avoiding disordered proteins when selecting targets to be solved using the commonly applied X-ray crystallography [32, 92].

We find that the most accurate disorder predictors according to CAID rely exclusively on the deep neural networks. More broadly, a recent analysis of the deep learning-based disorder predictors reveals that these methods rely on a wide range of network types and that they outperforms other types of predictive models [175]. The currently used deep network topologies include feed-forward (flDPnn [68]), convolutional (DeepCNF-D [176] and AUCpred [79]), recurrent (SPOT-Disorder [177], IDP-Seq2Seq [67], MetaPredict [71] and DeepCLD [72]) and convolutional/recurrent hybrids (SPOT-Disorder-Single, rawMSA [80], SPOT-Disorder2 [81] and RFPR-IDP [70]). Some of the recently released methods that predict specific types of functional disordered regions also use deep networks. Examples include SPOT-MoRF [178], MoRFPred_en [179] and en_DCNNMoRF [180] that predict protein-binding regions; DeepDISObind [181] that predicts protein and nucleic acid binding regions; and DisoLipPred [182] that finds putative disordered lipid binding regions. One recent advancement in deep networks that is yet to penetrate into the mainstream of the disorder predictions are the transformer networks. These architectures arguably improve over the convolutional and recurrent designs by applying attention mechanisms and positional embeddings. The transformer networks were already used with success in several related problems including prediction of tertiary protein structure [183], protein-protein interactions [184] and protein-compound interactions [185].

Moreover, authors of the recent flDPnn method conclude that predictive quality can be improved by extending inputs of the deep networks to cover additional functional and structural characteristics that are derived from the protein sequence [68]. We believe that further advances in the disorder prediction could stem from hybrid designs that combine multiple network topologies and tailor their predictive inputs. These hybrid designs should aim to mimic the underlying diversity of disorder types/flavors [186, 187]. For example, IDRs are classified into native coils, native pre-molten globules and native molten globules, which stems from differences in their conformational space [188]. IDRs also vary in length and localization in the sequence including typically short regions that are located at the termini [189] vs. long regions that can cover the whole length of the sequence [190, 191]. Simultaneous tuning of the topologies and inputs should offer sufficient amount of flexibility to accurately model different flavors of disorder. Hybridizing these models together will provide a more complete approach to cover the multiplicity of disorder and consequently should lead to further improvement in the predictive performance.

Another interesting future direction would be to expand the current scope of the QA tools to cover more disorder predictors. The current version of QUARTERplus works with six disorder predictors (DISOPRED3, IUPred, PONDR VSL2, DisEMBL, GlobPlot, and SPOT-Disorder-Single) that exclude the top-performers from CAID, such as AUCpred, rawMSA, SPOT-Disorder2, and flDPnn. Adding QA scores to the results produced by the latter set of methods would make them more practical, further boosting confidence in the results that they produce.

Declaration of interest

The authors declare no conflicts of interest.

Funding

This research was funded in part by the National Science Foundation (grants 2125218 and 2146027) and the Robert J. Mattauch Endowment funds to L.K.

References

1. Lieutaud, P., et al., *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): p. e1259708.
2. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions.* Intrinsically Disordered Proteins: Dynamics, Binding, and Function, 2019: p. 1-34.
3. Habchi, J., et al., *Introducing protein intrinsic disorder.* Chem Rev, 2014. **114**(13): p. 6561-88.
4. Dunker, A.K., et al., *What's in a name? Why these proteins are intrinsically disordered.* Intrinsically Disordered Proteins, 2013. **1**(1): p. e24157.
5. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
6. Xue, B., A.K. Dunker, and V.N. Uversky, *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): p. 137-49.
7. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
8. Di Domenico, T., I. Walsh, and S.C. Tosatto, *Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database.* BMC Bioinformatics, 2013. **14 Suppl 7**: p. S3.
9. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.
10. Peng, Z., M.J. Mizianty, and L. Kurgan, *Genome-scale prediction of proteins with long intrinsically disordered regions.* Proteins, 2014. **82**(1): p. 145-58.
11. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.* Journal of Molecular Recognition, 2005. **18**(5): p. 343-384.
12. Liu, J., et al., *Intrinsic disorder in transcription factors.* Biochemistry, 2006. **45**(22): p. 6873-88.
13. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cell Mol Life Sci, 2014. **71**(8): p. 1477-504.

14. Babu, M.M., *The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease.* Biochem Soc Trans, 2016. **44**(5): p. 1185-1200.
15. Peng, Z.L., et al., *More than just tails: intrinsic disorder in histone proteins.* Molecular Biosystems, 2012. **8**(7): p. 1886-1901.
16. Staby, L., et al., *Eukaryotic transcription factors: paradigms of protein intrinsic disorder.* Biochem J, 2017. **474**(15): p. 2509-2532.
17. Zhou, J.H., S.W. Zhao, and A.K. Dunker, *Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation.* Journal of Molecular Biology, 2018. **430**(16): p. 2342-2359.
18. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.* J Mol Recognit, 2005. **18**(5): p. 343-384.
19. Tantos, A., K.H. Han, and P. Tompa, *Intrinsic disorder in cell signaling and gene transcription.* Mol Cell Endocrinol, 2012. **348**(2): p. 457-65.
20. Zhao, B., et al., *Intrinsic Disorder in Human RNA-Binding Proteins.* J Mol Biol, 2021. **433**(21): p. 167229.
21. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
22. Wu, Z., et al., *In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces.* FEBS Lett, 2015. **589**(19 Pt A): p. 2561-9.
23. Fuxreiter, M., et al., *Disordered proteinaceous machines.* Chem Rev, 2014. **114**(13): p. 6806-43.
24. Zhao, B., et al., *IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell.* Cell Mol Life Sci, 2020.
25. Meng, F., et al., *Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments.* Int J Mol Sci, 2015. **17**(1).
26. Kuechler, E.R., et al., *Distinct Features of Stress Granule Proteins Predict Localization in Membraneless Organelles.* J Mol Biol, 2020. **432**(7): p. 2349-2368.
27. Orti, F., et al., *Insight into membraneless organelles and their associated proteins: Drivers, Clients and Regulators.* Comput Struct Biotechnol J, 2021. **19**: p. 3964-3977.
28. Cuevas-Velazquez, C.L. and J.R. Dinneny, *Organization out of disorder: liquid-liquid phase separation in plants.* Curr Opin Plant Biol, 2018. **45**(Pt A): p. 68-74.
29. Hu, G., et al., *Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity.* Proteomics, 2018: p. e1800243.
30. Kulkarni, P. and V.N. Uversky, *Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome.* Proteomics, 2018. **18**(21-22).
31. Uversky, V.N., *Bringing Darkness to Light: Intrinsic Disorder as a Means to Dig into the Dark Proteome.* Proteomics, 2018. **18**(21-22): p. e1800352.
32. Oldfield, C.J., et al., *Utilization of protein intrinsic disorder knowledge in structural proteomics.* Biochim Biophys Acta, 2013. **1834**(2): p. 487-98.
33. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
34. Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins.* Nucleic Acids Res, 2016. **D1**: p. D219-D227.
35. Sickmeier, M., et al., *DisProt: the Database of Disordered Proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D786-93.
36. Vucetic, S., et al., *DisProt: a database of protein disorder.* Bioinformatics, 2005. **21**(1): p. 137-40.
37. Quaglia, F., et al., *DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation.* Nucleic Acids Res, 2022. **50**(D1): p. D480-D487.
38. Le Gall, T., et al., *Intrinsic disorder in the Protein Data Bank.* J Biomol Struct Dyn, 2007. **24**(4): p. 325-42.
39. Fukuchi, S., et al., *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.* Nucleic Acids Res, 2014. **42**(Database issue): p. D320-5.

40. Schad, E., et al., *DIBS: a repository of disordered binding sites mediating interactions with ordered proteins.* Bioinformatics, 2018. **34**(3): p. 535-537.
41. Miskei, M., C. Antal, and M. Fuxreiter, *FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies.* Nucleic Acids Res, 2017. **45**(D1): p. D228-D235.
42. Ficho, E., et al., *MFIB: a repository of protein complexes with mutual folding induced by binding.* Bioinformatics, 2017. **33**(22): p. 3682-3684.
43. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder.* Bioinformatics, 2015. **31**(2): p. 201-8.
44. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
45. Kurgan, L., et al., *On the Importance of Computational Biology and Bioinformatics to the Origins and Rapid Progression of the Intrinsically Disordered Proteins Field*, in *Biocomputing 2020*. 2020. p. 149-158.
46. Zhao, B. and L. Kurgan, *Surveying over 100 predictors of intrinsic disorder in proteins.* Expert Rev Proteomics, 2021: p. 1-11.
47. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods.* Mol Biosyst, 2012. **8**(1): p. 114-21.
48. Katuwawala, A. and L. Kurgan, *Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins.* Biomolecules, 2020. **10**(12).
49. Li, J., et al., *An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014.* Int J Mol Sci, 2015. **16**(10): p. 23446-62.
50. Liu, Y., X. Wang, and B. Liu, *A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction.* Brief Bioinform, 2019. **20**(1): p. 330-346.
51. Meng, F., V. Uversky, and L. Kurgan, *Computational Prediction of Intrinsic Disorder in Proteins.* Curr Protoc Protein Sci, 2017. **88**: p. 2 16 1-2 16 14.
52. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.
53. Peng, Z.L. and L. Kurgan, *Comprehensive comparative assessment of in-silico predictors of disordered regions.* Curr Protein Pept Sci, 2012. **13**(1): p. 6-18.
54. Necci, M., et al., *A comprehensive assessment of long intrinsic protein disorder from the DisProt database.* Bioinformatics, 2018. **34**(3): p. 445-452.
55. He, B., et al., *Predicting intrinsic disorder in proteins: an overview.* Cell Res, 2009. **19**(8): p. 929-49.
56. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions.* Brief Bioinform, 2020. **21**(5): p. 1509-1522.
57. Necci, M., et al., *Critical assessment of protein intrinsic disorder prediction.* Nat Methods, 2021. **18**(5): p. 472-481.
58. Jin, Y. and R.L. Dunbrack, Jr., *Assessment of disorder predictions in CASP6.* Proteins, 2005. **61 Suppl 7**: p. 167-75.
59. Bordoli, L., F. Kiefer, and T. Schwede, *Assessment of disorder predictions in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 129-36.
60. Noivirt-Brik, O., J. Prilusky, and J.L. Sussman, *Assessment of disorder predictions in CASP8.* Proteins, 2009. **77 Suppl 9**: p. 210-6.
61. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10.* Proteins, 2014. **82 Suppl 2**: p. 127-37.
62. Melamud, E. and J. Moult, *Evaluation of disorder predictions in CASP5.* Proteins, 2003. **53 Suppl 6**: p. 561-5.
63. Monastyrskyy, B., et al., *Evaluation of disorder predictions in CASP9.* Proteins, 2011. **79 Suppl 10**: p. 107-18.

64. Williams, R.J., *The conformation properties of proteins in solution.* Biol Rev Camb Philos Soc, 1979. **54**(4): p. 389-437.
65. Orlando, G., et al., *Prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics.* bioRxiv, 2020: p. 2020.05.25.115253.
66. Dass, R., F.A.A. Mulder, and J.T. Nielsen, *ODiNPred: comprehensive prediction of protein order and disorder.* Sci Rep, 2020. **10**(1): p. 14780.
67. Tang, Y.J., Y.H. Pang, and B. Liu, *IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning.* Bioinformatics, 2021. **36**(21): p. 5177-5186.
68. Hu, G., et al., *flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions.* Nat Commun, 2021. **12**(1): p. 4438.
69. Erdos, G., M. Pajkos, and Z. Dosztanyi, *IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation.* Nucleic Acids Res, 2021. **49**(W1): p. W297-W303.
70. Liu, Y., X. Wang, and B. Liu, *RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins.* Brief Bioinform, 2021. **22**(2): p. 2000-2011.
71. Emenecker, R.J., D. Griffith, and A.S. Holehouse, *Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure.* Biophys J, 2021. **120**(20): p. 4312-4319.
72. Fang, M., et al., *DeepCLD: An efficient sequence-based predictor of intrinsically disordered proteins.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021: p. 1-1.
73. Ishida, T. and K. Kinoshita, *PrDOS: prediction of disordered protein regions from amino acid sequence.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W460-4.
74. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
75. Mizianty, M.J., et al., *Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources.* Bioinformatics, 2010. **26**(18): p. i489-96.
76. Mizianty, M.J., Z. Peng, and L. Kurgan, *MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles.* Intrinsically Disord Proteins, 2013. **1**(1): p. e24428.
77. Mizianty, M.J., V. Uversky, and L. Kurgan, *Prediction of intrinsic disorder in proteins using MFDp2.* Methods Mol Biol, 2014. **1137**: p. 147-62.
78. Lang, B. and M.M. Babu, *A community effort to bring structure to disorder.* Nat Methods, 2021. **18**(5): p. 454-455.
79. Wang, S., J. Ma, and J. Xu, *AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields.* Bioinformatics, 2016. **32**(17): p. i672-i679.
80. Mirabello, C. and B. Wallner, *rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments.* PLoS One, 2019. **14**(8): p. e0220182.
81. Hanson, J., et al., *SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning.* Genomics Proteomics Bioinformatics, 2019. **17**(6): p. 645-656.
82. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
83. Wilson, C.J., W.-Y. Choy, and M. Karttunen, *AlphaFold2: A role for disordered protein prediction?* bioRxiv, 2021: p. 2021.09.27.461910.
84. Aderinwale, T., et al., *Real-Time Structure Search and Structure Classification for AlphaFold Protein Models.* bioRxiv, 2021: p. 2021.10.21.465371.
85. Dosztanyi, Z., et al., *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* J Mol Biol, 2005. **347**(4): p. 827-39.
86. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
87. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.

88. Jones, D.T. and J.J. Ward, *Prediction of disordered regions in proteins from position specific score matrices.* Proteins, 2003. **53 Suppl 6**: p. 573-8.

89. Ward, J.J., et al., *The DISOPRED server for the prediction of protein disorder.* Bioinformatics, 2004. **20**(13): p. 2138-9.

90. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**: p. 208.

91. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder.* Proteins, 2005. **61 Suppl 7**: p. 176-82.

92. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.* Structure, 2003. **11**(11): p. 1453-9.

93. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-9.

94. Hanson, J., K. Paliwal, and Y. Zhou, *Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures.* J Chem Inf Model, 2018. **58**(11): p. 2369-2376.

95. Herold, M. and K.H. Nierhaus, *Incorporation of six additional proteins to complete the assembly map of the 50 S subunit from Escherichia coli ribosomes.* J Biol Chem, 1987. **262**(18): p. 8826-33.

96. Worbs, M., R. Huber, and M.C. Wahl, *Crystal structure of ribosomal protein L4 shows RNA-binding sites for ribosome incorporation and feedback control of the S10 operon.* EMBO J, 2000. **19**(5): p. 807-18.

97. Timsit, Y., et al., *The role of disordered ribosomal protein extensions in the early steps of eubacterial 50 S ribosomal subunit assembly.* Int J Mol Sci, 2009. **10**(3): p. 817-34.

98. Ishida, T. and K. Kinoshita, *Prediction of disordered regions in proteins based on the meta approach.* Bioinformatics, 2008. **24**(11): p. 1344-8.

99. Cheng, J.L., M.J. Sweredoski, and P. Baldi, *Accurate prediction of protein disordered regions by mining protein structure data.* Data Mining and Knowledge Discovery, 2005. **11**(3): p. 213-222.

100. Shimizu, K., S. Hirose, and T. Noguchi, *POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix.* Bioinformatics, 2007. **23**(17): p. 2337-8.

101. McGuffin, L.J., *Intrinsic disorder prediction from the analysis of multiple protein fold recognition models.* Bioinformatics, 2008. **24**(16): p. 1798-804.

102. Kozlowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins.* BMC Bioinformatics, 2012. **13**: p. 111.

103. Hecker, J., J.Y. Yang, and J.L. Cheng, *Protein disorder prediction at multiple levels of sensitivity and specificity.* Bmc Genomics, 2008. **9**.

104. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.

105. Su, C.T., C.Y. Chen, and C.M. Hsu, *iPDA: integrated protein disorder analyzer.* Nucleic Acids Research, 2007. **35**: p. W465-W472.

106. Hirose, S., et al., *POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions.* Bioinformatics, 2007. **23**(16): p. 2046-53.

107. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines.* Nucleic Acids Research, 2006. **34**: p. W164-W168.

108. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.* Bioinformatics, 2005. **21**(16): p. 3369-76.

109. Huang, Y.J., T.B. Acton, and G.T. Montelione, *DisMeta: a meta server for construct design and optimization.* Methods Mol Biol, 2014. **1091**: p. 3-16.

110. Prilusky, J., et al., *FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded.* Bioinformatics, 2005. **21**(16): p. 3435-8.

111. Barik, A., et al., *DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server.* J Mol Biol, 2020. **432**(11): p. 3379-3387.

112. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.

113. Necci, M., et al., *MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins.* Bioinformatics, 2017. **33**(9): p. 1402-1404.

114. Fan, X. and L. Kurgan, *Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus.* J Biomol Struct Dyn, 2014. **32**(3): p. 448-64.

115. Dunker, A.K., et al., *Intrinsic protein disorder in complete genomes.* Genome Inform Ser Workshop Genome Inform, 2000. **11**: p. 161-71.

116. Wang, C., V.N. Uversky, and L. Kurgan, *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): p. 1486-98.

117. Pentony, M.M. and D.T. Jones, *Modularity of intrinsic disorder in the human proteome.* Proteins, 2010. **78**(1): p. 212-21.

118. Colak, R., et al., *Distinct types of disorder in the human proteome: functional implications for alternative splicing.* PLoS Comput Biol, 2013. **9**(4): p. e1003030.

119. Oldfield, C.J., et al., *Codon selection reduces GC content bias in nucleic acids encoding for intrinsically disordered proteins.* Cell Mol Life Sci, 2020. **77**(1): p. 149-160.

120. Peng, Z., V.N. Uversky, and L. Kurgan, *Genes Encoding Intrinsic Disorder in Eukaryota Have High GC Content.* Intrinsically Disordered Proteins, 2016. **4**(1): p. e1262225.

121. Oates, M.E., et al., *D(2)P(2): database of disordered protein predictions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D508-16.

122. Piovesan, D., et al., *MobiDB: intrinsically disordered proteins in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D361-D367.

123. Potenza, E., et al., *MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins.* Nucleic Acids Res, 2015. **43**(Database issue): p. D315-20.

124. Di Domenico, T., et al., *MobiDB: a comprehensive database of intrinsic protein disorder annotations.* Bioinformatics, 2012. **28**(15): p. 2080-2081.

125. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D471-D476.

126. Zhao, B., et al., *DescribePROT: database of amino acid-level protein structure and function predictions.* Nucleic Acids Res, 2021. **49**(D1): p. D298-D308.

127. Cilia, E., et al., *From protein sequence to dynamics and disorder with DynaMine.* Nat Commun, 2013. **4**: p. 2741.

128. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins.* Bioinformatics, 2009. **25**(20): p. 2745-6.

129. Piovesan, D., et al., *FELLS: fast estimator of latent local structure.* Bioinformatics, 2017. **33**(12): p. 1889-1891.

130. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures.* Comput Chem, 1994. **18**(3): p. 269-85.

131. Lewis, T.E., et al., *Gene3D: Extensive prediction of globular domains in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D435-D439.

132. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.

133. Ghalwash, M.F., A.K. Dunker, and Z. Obradovic, *Uncertainty analysis in protein disorder prediction.* Mol Biosyst, 2012. **8**(1): p. 381-91.

134. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.

135. Faraggi, E., et al., *Fast and Accurate Accessible Surface Area Prediction Without a Sequence Profile.* Prediction of Protein Secondary Structure, 2017. **1484**: p. 127-136.

136. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.

137. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of Molecular Biology, 1999. **292**(2): p. 195-202.

138. Buchan, D.W.A. and D.T. Jones, *The PSIPRED Protein Analysis Workbench: 20 years on.* Nucleic Acids Research, 2019. **47**(W1): p. W402-W407.

139. Oldfield, C.J., Z. Peng, and L. Kurgan, *Disordered RNA-Binding Region Prediction with DisoRDPbind.* Methods Mol Biol, 2020. **2106**: p. 225-239.

140. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.

141. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.

142. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016.

143. Zhang, J., S. Ghadermarzi, and L. Kurgan, *Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins.* Bioinformatics, 2020. **36**(18): p. 4729-4738.

144. Zhang, J. and L. Kurgan, *SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences.* Bioinformatics, 2019. **35**(14): p. i343-i353.

145. Yan, J. and L. Kurgan, *DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues.* Nucleic Acids Res, 2017. **45**(10): p. e84.

146. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences.* Bioinformatics, 2016. **32**(12): p. i341-i350.

147. Teufel, F., et al., *SignalP 6.0 predicts all five types of signal peptides using protein language models.* Nat Biotechnol, 2022.

148. Almagro Armenteros, J.J., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.

149. Monzon, A.M., et al., *CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state.* Database (Oxford), 2016. **2016**.

150. Dinkel, H., et al., *ELM 2016--data update and new functionality of the eukaryotic linear motif resource.* Nucleic Acids Res, 2016. **44**(D1): p. D294-300.

151. consortium, P.D.-K., *PDBe-KB: collaboratively defining the biological context of structural data.* Nucleic Acids Res, 2022. **50**(D1): p. D534-D542.

152. Meszaros, B., et al., *PhaSePro: the database of proteins driving liquid-liquid phase separation.* Nucleic Acids Res, 2020. **48**(D1): p. D360-D367.

153. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.

154. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.

155. Wu, Z., et al., *Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins*, in *6th International Conference on Artificial Intelligence and Soft Computing*. 2017: Zakopane, Poland. p. 722-732.

156. Hu, G., et al., *Quality assessment for the putative intrinsic disorder in proteins.* Bioinformatics, 2019. **35**(10): p. 1692-1700.

157. Wu, Z., et al. *Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins*. 2017. Cham: Springer International Publishing.

158. Wu, Z., et al., *Prediction of Intrinsic Disorder with Quality Assessment Using QUARTER.* Methods Mol Biol, 2020. **2165**: p. 83-101.

159. Katuwawala, A., et al., *QUARTERplus: Accurate disorder predictions integrated with interpretable residue-level quality assessment scores.* Comput Struct Biotechnol J, 2021. **19**: p. 2597-2606.

160. Hu, G., et al., *Untapped Potential of Disordered Proteins in Current Druggable Human Proteome.* Curr Drug Targets, 2016. **17**(10): p. 1198-205.

161. Hosoya, Y. and J. Ohkanda, *Intrinsically Disordered Proteins as Regulators of Transient Biological Processes and as Untapped Drug Targets.* Molecules, 2021. **26**(8).

162. Biesaga, M., M. Frigole-Vivas, and X. Salvatella, *Intrinsically disordered proteins and biomolecular condensates as drug targets.* Curr Opin Chem Biol, 2021. **62**: p. 90-100.

163. Ambadipudi, S. and M. Zweckstetter, *Targeting intrinsically disordered proteins in rational drug discovery.* Expert Opin Drug Discov, 2016. **11**(1): p. 65-77.

164. Kurgan, L., M. Li, and Y. Li, *The Methods and Tools for Intrinsic Disorder Prediction and their Application to Systems Medicine*, in *Systems Medicine*, O. Wolkenhauer, Editor. 2021, Academic Press: Oxford. p. 159-169.

165. Hanson, J., et al., *Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning.* J Comput Biol, 2020. **27**(5): p. 796-814.

166. Zhao, Z., Z. Peng, and J. Yang, *Improving Sequence-Based Prediction of Protein-Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method.* J Chem Inf Model, 2018. **58**(7): p. 1459-1468.

167. Chowdhury, S., J. Zhang, and L. Kurgan, *In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome.* Proteomics, 2018: p. e1800064.

168. Flot, M., et al., *StackSSSPred: A Stacking-Based Prediction of Supersecondary Structure from Sequence.* Methods Mol Biol, 2019. **1958**: p. 101-122.

169. Dou, Y., B. Yao, and C. Zhang, *Prediction of Protein Phosphorylation Sites by Integrating Secondary Structure Information and Other One-Dimensional Structural Properties.* Methods Mol Biol, 2017. **1484**: p. 265-274.

170. Ghadermarzi, S., et al., *Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins.* Front Genet, 2019. **10**: p. 1075.

171. Nicolau, N., Jr. and S. Giuliatti, *Modeling and molecular dynamics of the intrinsically disordered e7 proteins from high- and low-risk types of human papillomavirus.* J Mol Model, 2013. **19**(9): p. 4025-37.

172. Shi, X., J. Zheng, and T. Yan, *Computational redesign of human respiratory syncytial virus epitope as therapeutic peptide vaccines against pediatric pneumonia.* J Mol Model, 2018. **24**(4): p. 79.

173. Zhong, B., et al., *Rational design of cyclic peptides to disrupt TGF-Beta/SMAD7 signaling in heterotopic ossification.* J Mol Graph Model, 2017. **72**: p. 25-31.

174. Liu, X. and J. Chen, *Modulation of p53 Transactivation Domain Conformations by Ligand Binding and Cancer-Associated Mutations.* Pac Symp Biocomput, 2020. **25**: p. 195-206.

175. Zhao, B. and L. Kurgan, *Deep learning in prediction of intrinsic disorder in proteins.* Computational and Structural Biotechnology Journal, 2022. **20**: p. 1286-1294.

176. Wang, S., et al., *DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields.* Int J Mol Sci, 2015. **16**(8): p. 17315-30.

177. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks.* Bioinformatics, 2017. **33**(5): p. 685-692.

178. Hanson, J., et al., *Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning.* Bioinformatics, 2020. **36**(4): p. 1107-1113.

179. Fang, C., et al., *MoRFPred_en: Sequence-based prediction of MoRFs using an ensemble learning strategy.* J Bioinform Comput Biol, 2019. **17**(6): p. 1940015.

180. Fang, C., et al., *Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method.* J Bioinform Comput Biol, 2019. **17**(1): p. 1950004.

181. Zhang, F., et al., *DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning.* Brief Bioinform, 2022. **23**(1).

182. Katuwawala, A., B. Zhao, and L. Kurgan, *DisoLipPred: Accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning.* Bioinformatics, 2021.

183. Hong, Y., J. Lee, and J. Ko, *A-Prot: protein structure modeling using MSA transformer.* BMC Bioinformatics, 2022. **23**(1): p. 93.
184. Ieremie, I., R.M. Ewing, and M. Niranjan, *TransformerGO: Predicting protein-protein interactions by modelling the attention between sets of gene ontology terms.* Bioinformatics, 2022.
185. Chen, L., et al., *TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments.* Bioinformatics, 2020. **36**(16): p. 4406-4414.
186. Necci, M., D. Piovesan, and S.C. Tosatto, *Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe.* Protein Sci, 2016. **25**(12): p. 2164-2174.
187. Deiana, A., et al., *Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell.* PLoS One, 2019. **14**(8): p. e0217889.
188. Uversky, V.N., *Unusual biophysics of intrinsically disordered proteins.* Biochim Biophys Acta, 2013. **1834**(5): p. 932-51.
189. Uversky, V.N., *The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini.* FEBS Lett, 2013. **587**(13): p. 1891-901.
190. Nielsen, J.T. and F.A. Mulder, *There is Diversity in Disorder-"In all Chaos there is a Cosmos, in all Disorder a Secret Order".* Front Mol Biosci, 2016. **3**: p. 4.
191. Xue, B., et al., *CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions.* FEBS Lett, 2009. **583**(9): p. 1469-74.