Machine Learning for Intrinsic Disorder Prediction

Bi Zhao[1], Lukasz Kurgan[1*]

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

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

Abstract

Intrinsic disorder in proteins is manifested by regions that lack stable structure under physiological conditions. Proteins with disordered regions are common across all kingdoms of life, facilitate many essential cellular functions and contribute to dark proteomes. They are associated with a wide spectrum of human diseases and consequently are considered as potent drug targets. Disordered regions have unique sequence signatures, making them predictable from protein sequences. Computational disorder prediction is a vibrant research area with over 40 years of history, which heavily depends on machine learning (ML) algorithms and innovations, such as meta learning and deep learning. We summarize a comprehensive collection of 73 ML-based disorder predictors, detail several most successful methods and survey related resources that predict disorder and disorder functions. We detail historical trends in the development of disorder predictors, highlighting the shifting focus from traditional ML methods, to meta-predictors, and most recently to the deep neural networks. We introduce a wide range of useful resources that support disorder and disorder function predictions including databases, webservers, and methods that provide quality assessment of disorder predictions. The availability of these numerous high-quality methods and resources ensures that the computational disorder predictions will continue to make substantial impact in key areas of research including rational drug design, structural genomics, and medicine.

1 Introduction

Intrinsic disorder in proteins is manifested by presence of regions that lack stable structure under physiological conditions [1-3]. These intrinsically disordered regions (IDRs) are classified as native coils, native pre-molten globules and native molten globules, signifying the fact that they be disordered to a different degree [4, 5]. More generally, IDRs can be seen as ensembles of interchanging conformations, with some regions being more expanded (native coils and native pre-molten globules) and other being more compact (native molten globule). Recent bioinformatics studies suggest that intrinsically disordered proteins (IDPs), i.e., proteins with IDRs, are common across all kingdoms of life, with particularly high levels in eukaryotes [6-9], and are distributed across different cellular compartment [10, 11]. IDPs are engaged in numerous essential cellular functions that include molecular assembly and recognition, signal transduction, cell cycle regulation, chromosomal packing, transcription and translation, to name but a few [8, 12-22]. They also contribute to the dark proteomes, defined as the non-resolved parts of the protein structure space [23-25].

Dysfunction of IDPs is associated with a wide spectrum of human diseases [26-28]. Examples include the Alzheimer's and Parkinson's diseases, Down's syndrome, prion diseases and dementia [29]. Moreover, IDPs facilitate regulatory and signaling functions that frequently rely on molecular interactions and consequently their misregulation and misinteractions are linked to cancers, diabetes and cardiovascular diseases [30-33]. IDPs play also key roles in viral genomes [34-38]. Given the prevalence and importance of IDPs in context of human diseases, IDPs such as α-synuclein, tau, p53 and BRCA-1 are attractive drug targets [26]. Moreover, novel approaches towards drug rational discovery efforts that target IDPs are being developed [39-43], further highlighting importance of this class of proteins.

Experimentally annotated IDRs can be collected from several databases, such as DisProt [44], PDB [45], IDEAL [46], DIBS [47], and MFIB [48]. The DisProt resource also offers functional annotations for the IDRs. While these resources provide access to very valuable experimental data, their size is relatively small compared to the hundreds of millions of protein sequences that are currently included the UniProt database [49]. For instance, the recent version 8.3 of DisProt covers approximately 2 thousand proteins [44]. The substantial and rapidly growing annotation gap has motivated development of computational predictors of IDRs [50-60]. The ability of these methods to generate accurate predictions of intrinsic disorder stems from the fact that IDRs have distinct sequence signatures compared to structured/ordered regions [61]. For instance, disordered regions typically have high net charge, low hydrophobicity and are depleted in aromatic residues when compared to their structured counterparts [5, 62, 63]. Availability of these computational methods has made substantial impact on the intrinsic disorder field, driving a rapid increase in the research on IDPs and IDRs [64].

Figure 1A shows disorder annotations for the SIR3 protein from *Saccharomyces cerevisiae* (UniProt ID: P06701) collected from DisProt (DisProt ID: DP00533) [44]. This transcriptional repressor that facilitates modulation of chromatin structure includes one IDR (positions 216 to 549) that was shown to interact with proteins and DNA [65]. We use this protein to demonstrate disorder prediction generated flDPnn [66], one of the currently best methods that recently won the Critical Assessment of Intrinsic Protein Disorder (CAID) [67, 68]. The disorder predictions are done at the residue level, which means that they are produced for each amino acid in the input protein sequence. The prediction consists of a numeric propensity score (higher value denotes higher likelihood for disorder) and a binary value (disordered vs. structured). The binary prediction is typically generated from the putative propensities, where amino acids with propensities higher than a threshold are categorized as disordered and the remaining residues are predicted as structured. Figure 1B shows the putative propensity of disorder (black line) and binary annotation of disordered residues (green horizontal bar) produced by flDPnn; the corresponding threshold is denoted by the dashed horizontal line. We note that the location of the predicted disordered residues is in close agreement with the experimental data, demonstrating that disorder predictors can produce very accurate results.

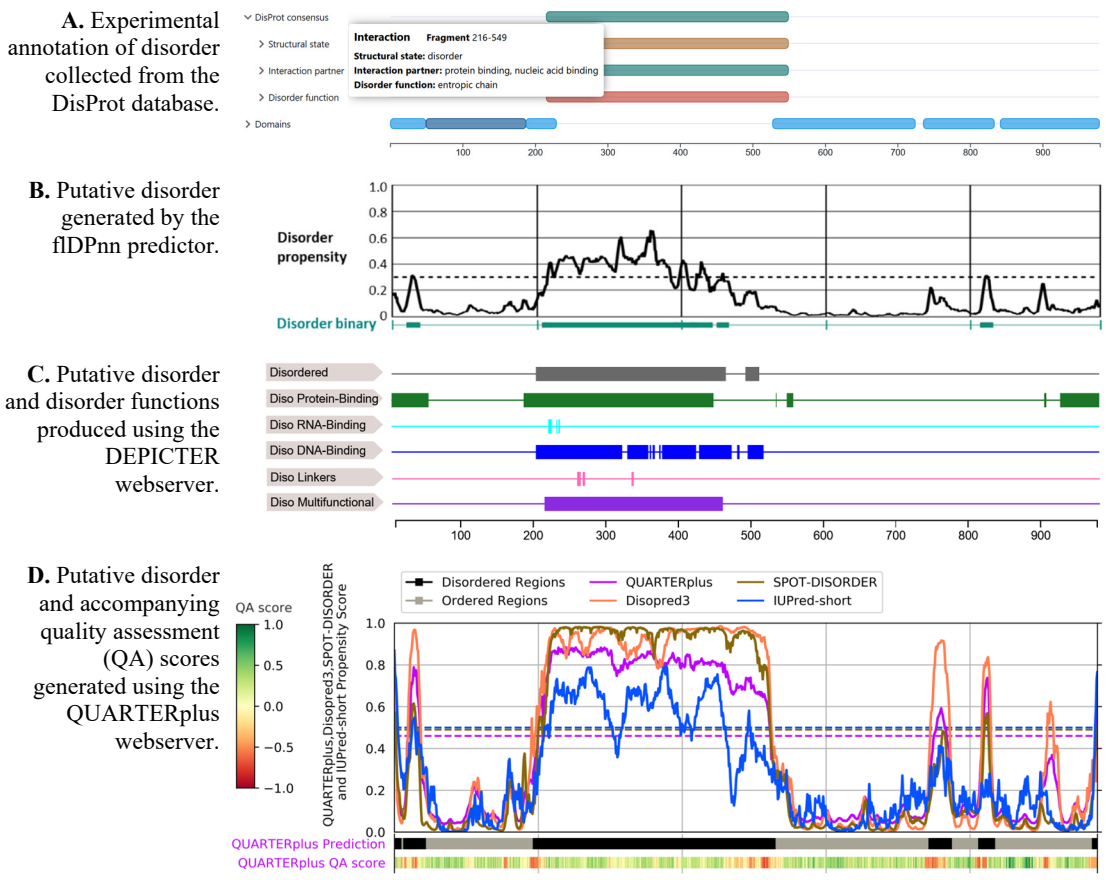

Figure 1. Experimental and predicted disorder annotations for the SIR3 protein (UniProt ID: P06701, DisProt ID: DP00533). Panel A shows the experimental annotation of the disordered regions (positions 216 to 549) collected from DisProt (https://www.disprot.org/) [44, 65]. Panel B shows prediction of disorder generated by the flDPnn method (http://biomine.cs.vcu.edu/servers/flDPnn/) [66] where black plot gives the putative numerical propensity of disorder and horizonal green bar corresponds to putative disordered regions derived from the propensity values. Panel C is the prediction of disorder and disorder functions produced by the DEPICTER webserver (http://biomine.cs.vcu.edu/servers/DEPICTER/) [69] that include disorder predictions using a consensus of the IUPred2A [70] and SPOT-Disorder-Single methods [71] (gray horizonal bar), putative disordered protein binding (green horizonal bar) by consensus of DisoRDPbind [72, 73], ANCHOR2 [70] and fMoRFpred [74], RNA binding (light blue horizonal bar) and DNA binding (dark blue horizonal bar) regions by DisoRDPbind [72, 73], disordered linkers (pink horizonal bar) by DFLpred [75], and putative multifunctional/moonlighting disordered regions identified by DMRpred (violet horizonal bar) [76]. Panel D gives the disorder prediction (black and gray horizontal bar) generated by the consensus of SPOT-DISORDER-Single [71], DISOPRED3 [77] and IUPred-short [78] that is accompanied by the quality assessment (QA) scores produced by QUARTERplus (http://biomine.cs.vcu.edu/servers/QUARTERplus/) [79]. The QA scores (color-coded horizonal bar) quantify quality of the consensus disorder prediction, i.e., residues identified with green and yellow colors are more likely to be accurately predicted compared to predictions colored in orange or red.

The flDPnn predictor and significant majority of other disorder predictors were developed using a machine learning (ML) approach [51, 54, 60]. This means that the developers of these methods used the available experimental disorder data to train predictive models using ML algorithms. Once trained and properly validated [68, 80], the resulting predictive models can be used to produce accurate predictions of disordered residues and regions for the millions of sequences that lack disorder annotations, like we illustrate in Figure 1B. This chapter

overviews disorder predictors that rely on the ML models. We produce a comprehensive list of these predictors and discuss the underlying ML algorithms used. Moreover, we provide a detailed description of several most successful methods. Finally, we briefly survey other related resources, including webservers and databases, that facilitate prediction of disorder and disorder functions.

2 Overview of disorder predictors

Over 100 disorder predictors have been developed to date [51, 54, 60]. They were reviewed in a about a dozen surveys [50-60]. The most recent review defines four distinct periods in the development of the disorder predictors [60]:

1. The ***first-generation*** predictors were developed between 1979 and 2001. The first ML-based method was developed by Romero, Obradovic, and Dunker in 1997 [81]. It relies on a shallow neural network that utilizes physical and chemical characteristics of the protein sequence as its inputs. Relatively few first-generation methods were developed.
2. The ***second-generation*** predictors date between 2002 and 2006. A significant event during this time period was the inclusion of disorder prediction assessment into the 5th Critical Assessment of Structure Prediction (CASP5) in 2003 [82]. This resulted in rapid popularization of this predictive area [50]. The second-generation predictors are typically based on somewhat simple predictive models, frequently relying on sequence scoring functions and shallow neural networks. The defining innovation was the use of evolutionary profiles produced from the position specific score matrix (PSSM) that is generated from the input protein sequences with the PSI-BLAST program [83, 84]. Representative second-generation methods include GlobPlot [85], IUPred [78, 86], PONDR predictors [87-90], DISOPRED [91], DisEMBL [92], and RONN [93].
3. The ***third-generation*** predictors were published between 2007 and 2015. One of the defining features of this time period was the introduction of meta-predictors, which generate disorder prediction by combining results produced by several disorder predictors. Popular third-generation meta-predictors include MFDp [94], CSpritz [95], PONDR-FIT [96] and DisCoP [97, 98]. We also note that the assessment of the disorder predictions continued biannually in the CASP7, CASP8, CASP9 and CASP10 experiments [80, 99-101], resulting in a steady stream of new predictors.
4. The ***fourth-generation*** period has started in 2016. This new generation of disorder predictors is defined by the introduction and development of deep learning models. Our analysis reveals that about half of the fourth-generation disorder predictors, 11 out of 23, rely on the deep neural networks [60]. Representative deep learning-based methods include AUCpred [102], SPOT-Disorder [103], SPOT-Disorder-Single [71], SPOT-Disorder2 [104], rawMSA [105] and flDPnn[66]. The focus on designing novel deep network-based predictors culminated in their convincing success in the most recent CAID community assessment [67, 68].

As the above historical overview suggests, disorder predictors utilize a broad spectrum of predictive models. They are typically divided into three categories based on their predictive models [51, 52, 54, 55, 58]: (1) sequence scoring function-based methods; (2) machine

learning approaches; and (3) meta-predictors. The ***sequence scoring*** function-based predictors use relatively simple additive and/or weighted functions, some of which are grounded in physical principles governing protein folding processes, to process information extracted from the input sequence and sequence-derived evolutionary information. Representative methods in this category include FoldIndex [106], IUPred [78, 86], IUPred2A [70] and IUPred3 [107]. The ***machine learning*** methods utilize sophisticated predictive models that are trained from experimental data using a variety of ML algorithms, such as support vector machine, conditional random field, random forest and a variety of neural networks. Well-known ML methods include DisEMBL [92], DISOPRED [91], PONDR [90], PrDOS [108], DISOPRED3 [77], fIDPnn [66], SPOT-Disorder2 [104], RawMSA [105] and AUCpred [102]. The ***meta-predictors*** utilize two or more disorder predictions as inputs to re-predict disorder to improve predictive performance when compared to the input predictions. Several empirical studies show that well-designed meta-predictors produce such improvements [97, 109-111]. Illustrative meta-predictors include metaPrDOS [112], MFDp [94], Cspritz [95], MFDp2 [113, 114], disCoP [97, 98] and MobiDB-lite [109]. Moreover, some meta-predictors use ML models to process inputs, which means that they belong to both categories. Corresponding examples include metaPrDOS [112] and MFDp [94].

3 Disorder prediction using machine learning

We focus on the machine learning disorder predictors. We searched for these methods using listing of methods that participated in community assessments [68, 80, 82, 99-101, 115] and a comprehensive selection of 13 previously published surveys, which also include comparative studies [50-55, 58, 60, 116-119]. This extensive search produced list of 73 ML-based disorder predictors that are summarized in Table 1. This Tables identifies when and where these methods were published and reviews ML models that they utilize.

Table 1. Summary of 73 predictors of intrinsic disorder that use machine learning models. The methods are sorted in the chronological order of their publication.

Disorder predictor	Year published	Ref.	Machine learning algorithms used
Predictor by Dunker *et al.*	1997	[81]	Shallow neural network
PONDR CaN-XT	1997	[120]	Shallow neural network
PONDR XL1	1997	[120]	Shallow neural network
PONDR VL-XT	2001	[89]	Shallow neural network
DisEMBL-REM465	2003	[92]	Shallow neural network
DisEMBL-HL	2003	[92]	Shallow neural network
DisEMBL-COIL	2003	[92]	Shallow neural network
DISOPRED	2003	[91]	Shallow neural network
DISOPRED2	2004	[121]	Support vector machine + Shallow neural network
DISpro	2005	[122, 123]	Shallow neural network
PONDR VL3	2005	[90]	Shallow neural network
PONDR VL3H	2005	[90]	Shallow neural network
PONDR VL3E	2005	[90]	Shallow neural network
RONN (JRONN)	2005	[93]	Shallow neural network
PONDR VSL1	2005	[87]	Regression
PROFbval	2006	[124]	Shallow neural network
PONDR VSL2B	2006	[87, 88]	Support vector machine

Disorder predictor	Year published	Ref.	Machine learning algorithms used
PONDR VSL2P	2006	[87, 88]	Support vector machine
Wiggle	2006	[125]	Support vector machine
Distill	2006	[126]	Shallow neural network
Spritz (Spritz3)	2006	[127]	Support vector machine
DisPSSMP	2006	[128]	Radial basis function networks
iPDA (DisPSSMP2)	2007	[129]	Radial basis function networks
POODLE-L	2007	[130]	Support vector machine
POODLE-S	2007	[131]	Support vector machine
POODLE-W	2007	[132]	Nearest neighbor
NORSnet	2007	[133]	Shallow neural network
PrDOS (PrDOS2)	2007	[108]	Support vector machine
Pdisorder	2007	N/A	Shallow neural network
UCON	2007	[133]	Shallow neural network
OnD-CRF (OnD-CRF2)	2008	[134]	Conditional random field
metaPrDOS (metaPrDOS2)	2008	[112]	Support vector machine
PreDisorder	2009	[135]	Shallow neural network
NN-CDF	2009	[136]	Shallow neural network
DRaai	2009	[137]	Random forest
MD	2009	[138]	Shallow neural network
UPforest-L	2009	[139]	Random forest
POODLE-I	2010	[140]	Support vector machine
MFDp	2010	[94]	Support vector machine
PONDR FIT	2010	[96]	Shallow neural network
Cspritz	2011	[95]	Shallow neural network
Espritz-D	2012	[141]	Shallow neural network
Espritz-N	2012	[141]	Shallow neural network
Espritz-X	2012	[141]	Shallow neural network
SPINE-D	2012	[142]	Shallow neural network
DNdisorder	2013	[143]	Deep neural network (restricted Boltzmann machine)
MFDp2	2013	[113, 114]	Support vector machine
disCoP	2014	[97, 98]	Regression
DynaMine	2014	[144, 145]	Regression
PON-Diso	2014	[146]	Random forest
s2D-1	2015	[147]	Shallow neural network
s2D-2	2015	[147]	Shallow neural network
DISOPRED3	2015	[77]	Support vector machine + Shallow neural network + Nearest neighbor
DisoMCS	2015	[148]	Conditional random field
DeepCNF-D	2015	[149]	Deep neural network (convolutional) + Conditional random field
AUCpred	2016	[102]	Deep neural network (convolutional) + Conditional random field
AUCpred-np	2016	[102]	Deep neural network (convolutional) + Conditional random field
DisPredict (DisPredict2)	2016	[150]	Support vector machine
SPOT-Disorder1	2017	[103]	Deep neural network (recurrent)
SPOT-Disorder-Single	2018	[71]	Deep neural network (hybrid: convolutional + recurrent)
Predictor by Zhao and Xue	2018	[151]	Decision tree + Shallow neural network
IDP-CRF	2018	[152]	Conditional random field
Spark-IDPP	2019	[153]	Support vector machine + Shallow neural network
IDP-FSP	2019	[154]	Conditional random field
rawMSA	2019	[105]	Deep neural network (hybrid: convolutional + recurrent)
SPOT-Disorder2	2019	[104]	Deep neural network (hybrid: convolutional + recurrent)
DisoMine	2020	N/A	Deep neural network (recurrent)
ODiNPred	2020	[155]	Shallow neural network
IDP-Seq2Seq	2020	[156]	Deep neural network (recurrent)

Disorder predictor	Year published	Ref.	Machine learning algorithms used
flDPnn	2021	[66]	Deep neural network (feed-forward)
flDPlr	2021	[66]	Regression
RFPR-IDP	2021	[107]	Deep neural network (hybrid: convolutional + recurrent)
Metapredict	2021	[157]	Deep neural network (recurrent)

Table 1 reveals that the first ML predictor was developed in 1997. These methods were developed at a relatively steady pace over the subsequent years. On average, close to 3 methods were developed annually, with 11 predictors published in the last 3 years. Figure 2 illustrates these trends. We observe a sharp spike in the development efforts between 2005 and 2007 when 21 methods were released. We speculate that this was fueled by the inclusion of the disorder prediction into CASP5 and CASP6 experiments [82, 115], which correspondingly grew from 6 participating methods in CASP5 to 20 in CASP6. Figure 2 also highlight selected popular and/or well performing methods that were developed over the years (in chronological order): PONDR XL1 [120] (1997), DisEMBL [92] and DISOPRED [91] (2003), PONDR VSL2B [87, 88] (2006), PrDOS [108] (2007), MFDp [94] (2010), Espritz [141] (2012), DISOPRED3 [77] (2015), AUCpred [102] (2016), SPOT-Disorder-Single [71] (2018), SPOT-Disorder2 [104] and RawMSA [105] (2019), and flDPnn [66] (2021).

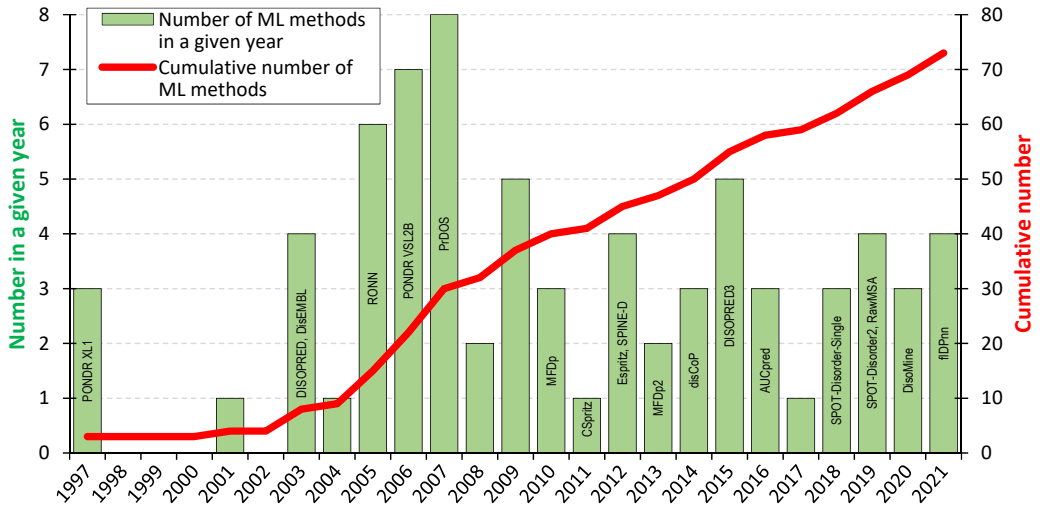

Figure 2. Timeline of the development of the machine learning-based predictors of intrinsic disorder. The green bars show the number of predictors developed in a given year. Selected popular and/or well-performing predictors are named inside the bars. The red line is the cumulative number of the predictors.

Table 1 shows that the 73 disorder predictors rely on a variety of different ML algorithms. Some of these predictors use an ensemble of models that were produced by multiple different ML algorithms. A case in point is DISOPRED3 that combines support vector machine, shallow neural network and nearest neighbor models [77]. We break down these algorithms in Figure 3. By far the most popular choices are shallow neural networks and support vector machines, which when combined account for 60% of all ML algorithms used. The shallow neural networks were particularly popular in early years, with nearly all predictors between 1997 and 2005 utilizing these models (Table 1). The support vector machines were commonly

used between 2006 and 2010, when 10 out of 25 predictors applied these models. Interestingly, the most popular ML models in recent years are deep neural networks, which were used by 6 out of 11 disorder predictors that were published since 2019. Deep neural networks differ from the shallow networks by inclusion of multiple hidden layers. The deep networks also often apply more advanced neuron types and more sophisticated architectures and connection patterns. The deep learning-based disorder predictors utilize a diverse collection of architecture types that cover feed-forward, recurrent, convolutional and hybrid topologies (Figure 3). The latter typically combine convolutional and recurrent topologies (Table 1).

The drive to use the deep networks partly stems from their popularity in related areas of protein bioinformatics [158, 159]. Some examples include prediction of protein function [160, 161], residue contacts [162-166], residue distances [167], binding residues [168-171], crystallization success [172], solvent accessibility [173, 174], secondary structure [174-176], posttranslational modification sites [177], and substrates and cleavage sites [178]. Another key factor contributing to the popularity of the deep learners is their success in recent assessments of disorder predictions [58, 68, 116]. For example, all of the best-performing methods that participated in the most recent CAID competition [67, 68], which include flDPnn [66], SPOT-Disorder2 [104], RawMSA [105] and AUCpred [102], rely on the deep networks. Interestingly, these methods use diverse network architectures including convolutional topology by AUCpred, a hybrid of convolutional and recurrent by SPOT-Disorder2 and RawMSA, and the feed-forward topology by flDPnn.

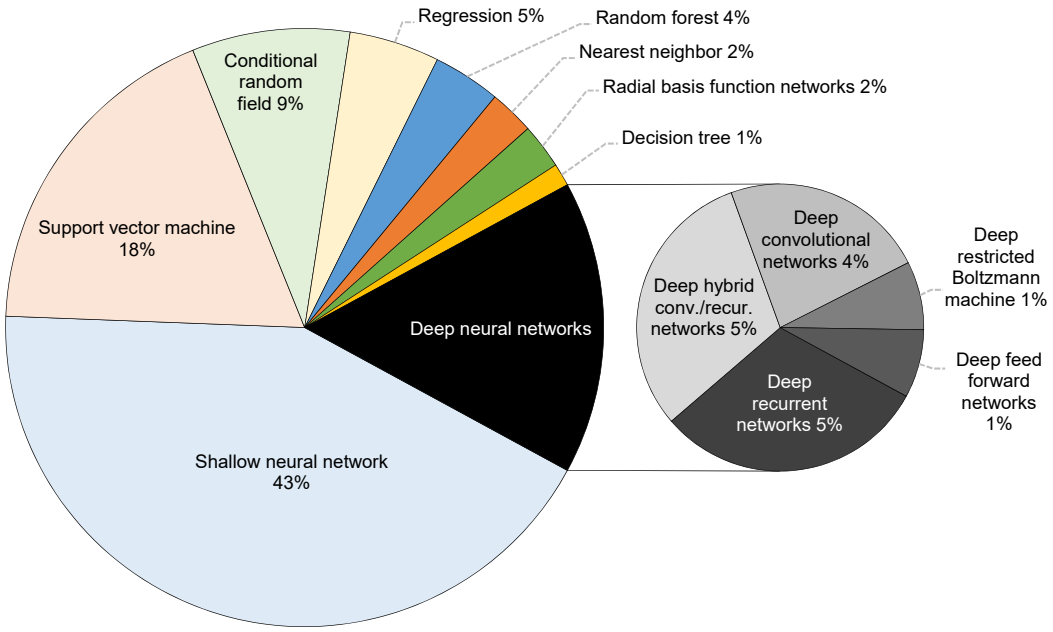

Figure 3. The breakdown of different types of machine learning models used by the disorder predictors. The smaller, gray-scale pie chart shows various types of the deep neural networks.

4 Selected machine learning-based disorder predictors

Several large-scale comparative studies that assessed predictive performance of the intrinsic disorder predictors were published in the past [68, 80, 82, 99-101, 115-119]. Among them, we highlight the community assessments where disorder predictors were evaluated on blind test datasets (i.e., datasets that were not available to the authors of predictors) by an independent group of assessors who do not take part in the competitions. This arguably ensures that the evaluation is fair across the participating predictors. The community assessments include (in chronological order) CASP5 [82], CASP6 [115], CASP7 [99], CASP8 [100], CASP9 [101], CASP10 [80] and CAID [68]. The largest and most recent assessments reported results for 28 disorder predictors in CASP10 [80] and 32 in CAID [68]. The top three methods in these two assessments utilize ML-based models. They include PrDOS [108], DISOPRED3 [77] and MFDp [94] from CASP10 [80]; and flDPnn [66]; SPOT-Disorder2 [104] and AUCpred [102] from CAID [66, 68]. Following, we provide detailed description of these six best-performing ML-based disorder predictors.

4.1 PrDOS

PrDOS was developed by Dr. Ishida at the University of Tokyo and Dr. Kinoshita at the Tohoku University. This method integrates two predictors: one based on an evolutionary profile of the protein sequence and the other based on the protein structure templates. The result is computed as a weighted average of the outputs from the two predictors. Comparative analysis reveals that this method outperformed all other disorder predictors in the CASP10 experiment [80].

Predictive model: Combination of SVM and the sequence alignments of sequences from PDB structures.

Citation data: 676 citations according to Google Scholar (January 2022); published in 2007 [108].

Availability: Webserver at https://prdos.hgc.jp/cgi-bin/top.cgi.

4.2 MFDp

The MFDp predictor was designed by Dr. Kurgan's team, which currently resides at the Virginia Commonwealth University. This is a meta-predictor that combines putative disorder predicted by four complementary methods: IUPred-long [78], IUPred-short [78], DISOPRED2 [121] and DISOclust [179]. The main innovation behind MFDp is the design of the consensus that relies on three SVM models that predict different sizes of IDRs (long, short, and all-size). The subsequently version of this predictor, MFDp2 [113], calibrates the outputs from the original MFDp using disorder content predictions generated by DisCon [180]. MFDp was ranked third in CASP10 [80], and did not participate in CAID.

Predictive model: Ensemble of three SVMs.

Citation data: 183 citations according to Google Scholar (January 2022); published in 2010 [94].

Availability: MFDp webserver at http://biomine.cs.vcu.edu/servers/MFDp/; MFDp2 webserver at http://biomine.cs.vcu.edu/servers/MFDp2/.

4.3 DISOPRED3

DISOPRED3 was released by Dr. Jones's lab at the University College London. The two prior versions of this popular disorder predictor were released in 2003 (DISOPRED) and 2004 (DISOPRED2) [91, 121]. DISOPRED3 integrates its own disorder predictor with outputs generated by DISOPRED2 and two ML-based predictions trained using long IDRs. This method also identifies protein binding sites in the putative IDRs using an additional SVM-based model. DISOPRED3 is a part of the PSIPRED workbench [181]. DISOPRED3 was ranked second in CASP10 and among top ten predictors in CAID [68, 80].

Predictive model: SVM, shallow neural network and nearest neighbor for disorder prediction. SVM for prediction of the disordered protein binding.

Citation data: 565 citations according to Google Scholar (January 2022); published in 2015 [77].

Availability: Webserver at http://bioinf.cs.ucl.ac.uk/psipred/; standalone code at http://bioinfadmin.cs.ucl.ac.uk/downloads/DISOPRED/.

4.4 AUCpred

AUCpred was created by Dr. Xu's group at the Toyota Technological Institute in Chicago in collaboration with researchers from the University of Chicago. One of the innovations was that the underlying predictive model that relies on deep convolutional network was trained to maximize area under the ROC curve (AUC). This method offers two options: prediction with or without the use of the evolutionary profile. The latter prediction is much faster (10 seconds for an average length protein chain) and slightly less accurate than the outputs of the former design. AUCpred was ranked among the top three methods in the CAID experiment [66, 68].

Predictive model: Deep convolutional neural network combined with conditional random fields.

Citation data: 66 citations according to Google Scholar (January 2022); published in 2016 [102].

Availability: Webserver at http://raptorx.uchicago.edu/StructurePropertyPred/predict/.

4.5 SPOT-Disorder2

The SPOT-Disorder2 method was released by Prof. Zhou's lab, which is now located at the Shenzhen Bay Laboratory. The earlier SPOT-Disorder version of this method was published in 2017 [103]. SPOT-Disorder2 improves over SPOT-Disorder by integrating several network topologies rather than using one long short-term memory bidirectional recurrent neural network (LSTM-BRNN). The second version utilizes evolutionary profile and predictions from SPOT-1D method [182] as inputs. Besides predicting disorder, SPOT-Disorder2 outputs putative semi-disordered regions which can be used to identify molecular recognition features (MoRFs) [74, 183]. CAID results place this method among the top three predictors, however, SPOT-Disorder2 suffers long runtime when compared to its close challengers [66, 68].

Predictive model: Hybrid deep neural network that combines convolutional and recurrent topologies.

Citation data: 42 citations according to Google Scholar (January 2022); published in 2019 [104].

Availability: Webserver at https://sparks-lab.org/server/spot-disorder2/; standalone code at https://sparks-lab.org/downloads/.

4.6 flDPnn

The flDPnn predictor was developed in Prof. Kurgan's lab, which is currently located at the Virginia Commonwealth University, in collaboration with researchers from the Nankai University. The defining features of this method include innovative predictive inputs that incorporate extended sequences profile and protein-level feature encoding, ability to predict selected functions for the putative IDRs that it predicts, and low runtime. The functions covered by flDPnn include disordered linkers and interaction with proteins, DNA and RNA. FlDPnn produces the disorder and disorder function predictions very quickly, in about 5 seconds for an average size protein. Comparative analysis in CAID reveals that flDPnn ranks among the top three predictors in that experiment and that it is at least an order of magnitude faster than these competitors [66, 68].

Predictive model: Deep feed-forward neural network for the disorder prediction. Ensemble of four random forest models for the prediction of disorder functions.

Citation data: 6 citations according to Google Scholar (January 2022); published in 2021 [66].

Availability: Webserver at http://biomine.cs.vcu.edu/servers/flDPnn/; standalone code at https://gitlab.com/sina.ghadermarzi/fldpnn.

5 Related resources

Nowadays, users are provided with access to webservers and implementations for many disorder predictors. However, making these predictions could be inconvenient, particularly in scenarios where users would like to secure multiple disorder predictions for the same protein or predictions for a large number of proteins. A convenient alternative to making prediction using individual methods is to collect pre-computed predictions from one of the currently available databases: D^2P^2 (Database of Disorder Protein Predictions; https://d2p2.pro/) [184], MobiDB (https://mobidb.bio.unipd.it/) [185, 186], and DescribePROT (http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/) [187]. Each of these three databases provides instantaneous access to results generated by several disorder predictors for large datasets of proteins ranging from 1.35 million proteins from 83 genomes in DescribePROT, 10.43 million proteins from 1,765 genomes in D^2P^2, to 219.74 million proteins in MobiDB. One of the key features of the MobiDB resource is the inclusion of the consensus disorder prediction produced by MobiDB-lite method [109] as well as curated experimental annotations of disorder that are collected from several source including DisProt [44], IDEAL [46], ELM [188], MFIB [48], DIBS [47], FuzDB [189] and PhasePro [190]. While D^2P^2 and

MobiDB primarily focus on the disorder predictions, DescribePROT also provides predictions of other structural and functional characteristics of proteins. These include putative solvent accessibility predicted by ASAquick [191], putative disordered linked by DFLpred [75], putative protein-binding residues by DisoRDPbind [72, 73, 192], MoRFChibi [193] and SCRIBER [194, 195], putative DNA-binding and RNA-binding residues by DisoRDPbind and DRNApred [196], secondary structure by PSIPRED [197], signal peptides by SignalP [198, 199], disorder by PONDR VSL2B [87, 88], and alignment profiles produced by MMseqs2 [200, 201]. In total, the most recent release 1.4 of DescribePROT provides access to over 7.8 billion residue-level predictions.

However, we note that users must still rely on disorder predictors when they want to predict sequences that are not included in a given database. A very useful resource in that case is DEPICTER (DisorderEd PredictIon CenTER; http://biomine.cs.vcu.edu/servers/DEPICTER/) [69]. This unique webserver generates a comprehensive collection of disorder and disorder function predictions. It provides consensus disorder predictions using results produced by the fast IUPred-short [78], IUPred-long [78] and SPOT-Disorder-Single [71] methods. These predictions are accompanied by the disordered linker predictions made by DFLpred [75], putative disordered regions that interact with proteins and nucleic acids that are predicted by combining results of fMoRFpred [74], DisoRDPbind [72] and ANCHOR2 [70], and putative disordered multifunctional (moonlighting) regions generated by DMRpred [76]. Figure 1C shows results computed by DEPICTER for the SIR3 protein (UniProt ID: P06701; DisProt ID: DP00533). DEPICTER suggests that the putative IDRs (grey horizontal bar) is multifunctional (violet horizontal bar) and that it binds DNA (dark blue horizontal bar) and proteins (green horizontal bar). These predictions are in good agreement with the experimental disorder annotations shown in Figure 1A.

Another recent advance is the development of methods that provide interpretable residue-level quality assessment scores [202]: QUARTER (http://biomine.cs.vcu.edu/servers/QUARTER/) [203, 204] and QUARTERplus (http://biomine.cs.vcu.edu/servers/QUARTERplus/) [79]. The scores produced by these tools can be used to identify regions where the quality of the disorder predictions generated by several popular methods, such as DISOPRED3, IUPred, PONDR VSL2B and disEMBL, is high. QUARTERplus relies on a deep convolutional neural network to make accurate, consensus-based disorder predictions accompanied by the quality assessment scores, which allow the users to easily pinpoint which disorder predictions are more trustworthy. We illustrate this in Figure 1D where the disorder predictions shown using the black and gray horizontal bar are annotated with the quality assessment scores, i.e., color-coded horizontal bar where residues identified with green and yellow colors are more likely to be accurately predicted compared to predictions colored in orange or red. We note that the short putative IDRs that were identified at both sequence termini are marked in red/orange, which suggests that these predictions are likely incorrect. In contrast, the long putative IDRs in the middle of the sequence is marked in yellow/green, suggesting that this disorder prediction is likely accurate. These color-coded annotations concur with the experimental disorder annotations from Figure 1A, signifying the usefulness of the quality assessment scores that QUARTERplus produces.

We also briefly overview computational predictors of disorder functions. There are well over a dozen predictors of disordered protein-binding regions [205], including recent methods, such as OPAL+ [206], MoRFPred_en [207], FLIPPER [208] and SPOT-MoRF [209]. Moreover, users can utilize DisoRDPbind [72, 73, 192] and DeepDISOBind [169] to identify putative IDRs that interact with DNA, RNA and proteins, as well as DisoLipPred [170] that predicts IDRs that bind lipids. There are also two methods that predict disordered linker regions, DFLpred [75] and IPOD [210]. We stress the fact that prediction of binding IDRs (i.e, IDRs that bind proteins, DNA, RNA and small ligands) was recently assessed in the CAID experiment [68]. CAID found that ANCHOR2 [70], DisoRDPbind [72, 73, 192] and MoRFCHiBi [193] are the most accurate predictors of binding IDRs. However, this assessment concluded that "*disordered binding regions remain hard to predict*", suggesting that disorder function predictors should be further improved [68]. Interested readers can find more details in several recent surveys on this topic [51, 205, 211].

6 Summary

Disorder prediction is a vibrant and very active research area that heavily relies on ML models and innovations, including meta learning and deep learning. We identified 73 ML-based disorder predictors that were developed in the last four decades. We found that the original focus on traditional ML methods, such as shallow neural networks and support vector machines, which dominated this field until mid-2000s has shifted towards the meta-predictors in late 2010s. This was subsequently followed by a transition to the deep neural networks in around 2015. Given the success of deep learners in the recent CAID experiment [66, 68] and their popularity in the broader protein bioinformatics area [158, 159], we anticipate that the development of deep neural network-based disorder predictors will continue in a near future. We also stress the availability of many useful resources that support disorder and disorder function predictions including databases, such as D^2P^2 [184], MobiDB [185, 186], and DescribePROT [187], comprehensive webservers, such as DEPICTER [69], and methods that provide quality assessment of disorder predictions, such as QUARTERplus [79]. The easy access to these numerous methods and resources ensures that the computational disorder predictions will continue to make substantial impact in other key areas of research, such as rational drug design [42, 43, 212-214], structural genomics [24, 92, 215], study of human diseases [216], and systems medicine [59].

Funding

This work 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. Dunker, A.K., et al., *What's in a name? Why these proteins are intrinsically disordered.* Intrinsically Disordered Proteins, 2013. **1**(1): p. e24157.

2. 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.
3. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions.* Intrinsically Disordered Proteins: Dynamics, Binding, and Function, 2019: p. 1-34.
4. Uversky, V.N., *Unusual biophysics of intrinsically disordered proteins.* Biochim Biophys Acta, 2013. **1834**(5): p. 932-51.
5. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions*, in *Intrinsically Disordered Proteins*, N. Salvi, Editor. 2019, Academic Press. p. 1-34.
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. 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.
9. 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.
10. Zhao, B., et al., *IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell.* Cell Mol Life Sci, 2020.
11. 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).
12. 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.
13. Liu, J., et al., *Intrinsic disorder in transcription factors.* Biochemistry, 2006. **45**(22): p. 6873-88.
14. 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.
15. 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.
16. Peng, Z.L., et al., *More than just tails: intrinsic disorder in histone proteins.* Molecular Biosystems, 2012. **8**(7): p. 1886-1901.
17. Xie, H.B., et al., *Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.* Journal of Proteome Research, 2007. **6**(5): p. 1917-1932.
18. Xie, H.B., et al., *Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.* Journal of Proteome Research, 2007. **6**(5): p. 1882-1898.
19. Staby, L., et al., *Eukaryotic transcription factors: paradigms of protein intrinsic disorder.* Biochem J, 2017. **474**(15): p. 2509-2532.
20. 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).
21. Na, I., et al., *Autophagy-related intrinsically disordered proteins in intra-nuclear compartments.* Mol Biosyst, 2016. **12**(9): p. 2798-817.
22. Peng, Z., et al., *Resilience of death: intrinsic disorder in proteins involved in the programmed cell death.* Cell Death Differ, 2013. **20**(9): p. 1257-67.
23. Bhowmick, A., et al., *Finding Our Way in the Dark Proteome.* J Am Chem Soc, 2016. **138**(31): p. 9730-42.

24. 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.
25. Kulkarni, P. and V.N. Uversky, *Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome.* Proteomics, 2018. **18**(21-22).
26. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically disordered proteins in human diseases: introducing the D2 concept.* Annu Rev Biophys, 2008. **37**: p. 215-46.
27. Midic, U., et al., *Protein disorder in the human diseasome: unfoldomics of human genetic diseases.* BMC Genomics, 2009. **10 Suppl 1**: p. S12.
28. Xie, H., et al., *Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.* J Proteome Res, 2007. **6**(5): p. 1917-32.
29. Uversky, V.N., *The triple power of D(3): protein intrinsic disorder in degenerative diseases.* Front Biosci (Landmark Ed), 2014. **19**: p. 181-258.
30. Uversky, V.N., et al., *Unfoldomics of human diseases: linking protein intrinsic disorder with diseases.* BMC Genomics, 2009. **10 Suppl 1**: p. S7.
31. Cheng, Y., et al., *Abundance of intrinsic disorder in protein associated with cardiovascular disease.* Biochemistry, 2006. **45**(35): p. 10448-60.
32. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
33. Al-Jiffri, O.H., et al., *Intrinsic disorder in biomarkers of insulin resistance, hypoadiponectinemia, and endothelial dysfunction among the type 2 diabetic patients.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1171278.
34. Fan, X., et al., *The intrinsic disorder status of the human hepatitis C virus proteome.* Mol Biosyst, 2014. **10**(6): p. 1345-63.
35. Xue, B., et al., *Protein intrinsic disorder as a flexible armor and a weapon of HIV-1.* Cell Mol Life Sci, 2012. **69**(8): p. 1211-59.
36. Xue, B. and V.N. Uversky, *Intrinsic disorder in proteins involved in the innate antiviral immunity: another flexible side of a molecular arms race.* J Mol Biol, 2014. **426**(6): p. 1322-50.
37. Alshehri, M.A., et al., *On the Prevalence and Potential Functionality of an Intrinsic Disorder in the MERS-CoV Proteome.* Viruses, 2021. **13**(2).
38. Xue, B., et al., *Structural disorder in viral proteins.* Chem Rev, 2014. **114**(13): p. 6880-911.
39. Cheng, Y., et al., *Rational drug design via intrinsically disordered protein.* Trends Biotechnol, 2006. **24**(10): p. 435-42.
40. Dunker, A.K. and V.N. Uversky, *Drugs for 'protein clouds': targeting intrinsically disordered transcription factors.* Curr Opin Pharmacol, 2010. **10**(6): p. 782-8.
41. Ghadermarzi, S., et al., *Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins.* Front Genet, 2019. **10**: p. 1075.
42. Hu, G., et al., *Untapped Potential of Disordered Proteins in Current Druggable Human Proteome.* Curr Drug Targets, 2016. **17**(10): p. 1198-205.
43. Uversky, V.N., *Intrinsically disordered proteins and novel strategies for drug discovery.* Expert Opin Drug Discov, 2012. **7**(6): p. 475-88.
44. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
45. Le Gall, T., et al., *Intrinsic disorder in the Protein Data Bank.* J Biomol Struct Dyn, 2007. **24**(4): p. 325-42.

46. Fukuchi, S., et al., *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.* Nucleic Acids Research, 2014. **42**(D1): p. D320-D325.
47. Schad, E., et al., *DIBS: a repository of disordered binding sites mediating interactions with ordered proteins.* Bioinformatics, 2018. **34**(3): p. 535-537.
48. Ficho, E., et al., *MFIB: a repository of protein complexes with mutual folding induced by binding.* Bioinformatics, 2017. **33**(22): p. 3682-3684.
49. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.
50. He, B., et al., *Predicting intrinsic disorder in proteins: an overview.* Cell Res, 2009. **19**(8): p. 929-49.
51. 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.
52. 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.
53. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods.* Mol Biosyst, 2012. **8**(1): p. 114-21.
54. 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.
55. 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.
56. Pryor, E.E., Jr. and M.C. Wiener, *A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder.* Biophys J, 2014. **106**(8): p. 1638-49.
57. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins.* Brief Bioinform, 2010. **11**(2): p. 225-43.
58. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions.* Brief Bioinform, 2020. **21**(5): p. 1509-1522.
59. 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.
60. Zhao, B. and L. Kurgan, *Surveying over 100 predictors of intrinsic disorder in proteins.* Expert Rev Proteomics, 2021: p. 1-11.
61. Campen, A., et al., *TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder.* Protein Pept Lett, 2008. **15**(9): p. 956-63.
62. Uversky, V.N., *A decade and a half of protein intrinsic disorder: biology still waits for physics.* Protein Sci, 2013. **22**(6): p. 693-724.
63. Yan, J., et al., *Structural and functional analysis of "non-smelly" proteins.* Cell Mol Life Sci, 2020. **77**(12): p. 2423-2440.
64. 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.
65. McBryant, S.J., C. Krause, and J.C. Hansen, *Domain organization and quaternary structure of the Saccharomyces cerevisiae silent information regulator 3 protein, Sir3p.* Biochemistry, 2006. **45**(51): p. 15941-8.
66. Hu, G., et al., *flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions.* Nat Commun, 2021. **12**(1): p. 4438.

67. Lang, B. and M.M. Babu, *A community effort to bring structure to disorder.* Nat Methods, 2021. **18**(5): p. 454-455.
68. Necci, M., et al., *Critical assessment of protein intrinsic disorder prediction.* Nat Methods, 2021. **18**(5): p. 472-481.
69. Barik, A., et al., *DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server.* J Mol Biol, 2020. **432**(11): p. 3379-3387.
70. 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.
71. 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.
72. 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.
73. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.
74. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.
75. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences.* Bioinformatics, 2016. **32**(12): p. i341-i350.
76. Meng, F. and L. Kurgan, *High-throughput prediction of disordered moonlighting regions in protein sequences.* Proteins, 2018. **86**(10): p. 1097-1110.
77. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
78. 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.
79. 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.
80. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10.* Proteins, 2014. **82 Suppl 2**: p. 127-37.
81. Romero, P., et al., *Identifying disordered regions in proteins from amino acid sequence.* 1997 Ieee International Conference on Neural Networks, Vols 1-4, 1997: p. 90-95.
82. Melamud, E. and J. Moult, *Evaluation of disorder predictions in CASP5.* Proteins, 2003. **53 Suppl 6**: p. 561-5.
83. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
84. Hu, G. and L. Kurgan, *Sequence Similarity Searching.* Curr Protoc Protein Sci, 2019. **95**(1): p. e71.
85. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.
86. 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.
87. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder.* Proteins, 2005. **61 Suppl 7**: p. 176-82.
88. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**: p. 208.

89. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.
90. Peng, K., et al., *Optimizing long intrinsic disorder predictors with protein evolutionary information.* J Bioinform Comput Biol, 2005. **3**(1): p. 35-60.
91. 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.
92. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.* Structure, 2003. **11**(11): p. 1453-9.
93. 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.
94. 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.
95. Walsh, I., et al., *CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W190-6.
96. Xue, B., et al., *PONDR-FIT: a meta-predictor of intrinsically disordered amino acids.* Biochim Biophys Acta, 2010. **1804**(4): p. 996-1010.
97. 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.
98. Oldfield, C.J., et al., *Computational Prediction of Intrinsic Disorder in Protein Sequences with the disCoP Meta-predictor.* Methods Mol Biol, 2020. **2141**: p. 21-35.
99. Bordoli, L., F. Kiefer, and T. Schwede, *Assessment of disorder predictions in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 129-36.
100. Noivirt-Brik, O., J. Prilusky, and J.L. Sussman, *Assessment of disorder predictions in CASP8.* Proteins, 2009. **77 Suppl 9**: p. 210-6.
101. Monastyrskyy, B., et al., *Evaluation of disorder predictions in CASP9.* Proteins, 2011. **79 Suppl 10**: p. 107-18.
102. 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.
103. 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.
104. Hanson, J., et al., *SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning.* Genomics Proteomics Bioinformatics, 2019. **17**(6): p. 645-656.
105. Mirabello, C. and B. Wallner, *rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments.* PLoS One, 2019. **14**(8): p. e0220182.
106. 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.
107. 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.
108. 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.
109. Necci, M., et al., *MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins.* Bioinformatics, 2017. **33**(9): p. 1402-1404.
110. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.
111. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *DISOselect: Disorder predictor selection at the protein level.* Protein Sci, 2020. **29**(1): p. 184-200.

112. Ishida, T. and K. Kinoshita, *Prediction of disordered regions in proteins based on the meta approach.* Bioinformatics, 2008. **24**(11): p. 1344-8.
113. 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.
114. Mizianty, M.J., V. Uversky, and L. Kurgan, *Prediction of intrinsic disorder in proteins using MFDp2.* Methods Mol Biol, 2014. **1137**: p. 147-62.
115. Jin, Y. and R.L. Dunbrack, Jr., *Assessment of disorder predictions in CASP6.* Proteins, 2005. **61 Suppl 7**: p. 167-75.
116. 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).
117. 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.
118. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder.* Bioinformatics, 2015. **31**(2): p. 201-8.
119. Necci, M., et al., *A comprehensive assessment of long intrinsic protein disorder from the DisProt database.* Bioinformatics, 2018. **34**(3): p. 445-452.
120. Romero, Obradovic, and K. Dunker, *Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family.* Genome Inform Ser Workshop Genome Inform, 1997. **8**: p. 110-124.
121. Ward, J.J., et al., *The DISOPRED server for the prediction of protein disorder.* Bioinformatics, 2004. **20**(13): p. 2138-9.
122. Hecker, J., J.Y. Yang, and J.L. Cheng, *Protein disorder prediction at multiple levels of sensitivity and specificity.* Bmc Genomics, 2008. **9**.
123. 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.
124. Schlessinger, A., G. Yachdav, and B. Rost, *PROFbval: predict flexible and rigid residues in proteins.* Bioinformatics, 2006. **22**(7): p. 891-3.
125. Gu, J., M. Gribskov, and P.E. Bourne, *Wiggle - Predicting functionally flexible regions from primary sequence.* Plos Computational Biology, 2006. **2**(7): p. 769-785.
126. Bau, D., et al., *Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins.* BMC Bioinformatics, 2006. **7**: p. 402.
127. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W164-8.
128. Su, C.T., C.Y. Chen, and Y.Y. Ou, *Protein disorder prediction by condensed PSSM considering propensity for order or disorder.* BMC Bioinformatics, 2006. **7**: p. 319.
129. Su, C.T., C.Y. Chen, and C.M. Hsu, *iPDA: integrated protein disorder analyzer.* Nucleic Acids Research, 2007. **35**: p. W465-W472.
130. 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.
131. 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. 2337-2338.
132. Shimizu, K., et al., *Predicting mostly disordered proteins by using structure-unknown protein data.* Bmc Bioinformatics, 2007. **8**.
133. Schlessinger, A., M. Punta, and B. Rost, *Natively unstructured regions in proteins identified from contact predictions.* Bioinformatics, 2007. **23**(18): p. 2376-84.

134. Wang, L. and U.H. Sauer, *OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields.* Bioinformatics, 2008. **24**(11): p. 1401-2.
135. Deng, X., J. Eickholt, and J. Cheng, *PreDisorder: ab initio sequence-based prediction of protein disordered regions.* BMC Bioinformatics, 2009. **10**: p. 436.
136. 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.
137. Han, P.F., X.Z. Zhang, and Z.P. Feng, *Predicting disordered regions in proteins using the profiles of amino acid indices.* Bmc Bioinformatics, 2009. **10**.
138. Schlessinger, A., et al., *Improved disorder prediction by combination of orthogonal approaches.* PLoS One, 2009. **4**(2): p. e4433.
139. Han, P.F., et al., *Large-scale prediction of long disordered regions in proteins using random forests.* Bmc Bioinformatics, 2009. **10**.
140. Hirose, S., K. Shimizu, and T. Noguchi, *POODLE-I: disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach.* In Silico Biol, 2010. **10**(3): p. 185-91.
141. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-9.
142. Zhang, T., et al., *SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method.* J Biomol Struct Dyn, 2012. **29**(4): p. 799-813.
143. Eickholt, J. and J. Cheng, *DNdisorder: predicting protein disorder using boosting and deep networks.* BMC Bioinformatics, 2013. **14**: p. 88.
144. Cilia, E., et al., *From protein sequence to dynamics and disorder with DynaMine.* Nat Commun, 2013. **4**: p. 2741.
145. Cilia, E., et al., *The DynaMine webserver: predicting protein dynamics from sequence.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W264-70.
146. Ali, H., et al., *Performance of protein disorder prediction programs on amino acid substitutions.* Hum Mutat, 2014. **35**(7): p. 794-804.
147. Sormanni, P., et al., *The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins.* J Mol Biol, 2015. **427**(4): p. 982-996.
148. Wang, Z., et al., *DisoMCS: Accurately Predicting Protein Intrinsically Disordered Regions Using a Multi-Class Conservative Score Approach.* PLoS One, 2015. **10**(6): p. e0128334.
149. 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.
150. Iqbal, S. and M.T. Hoque, *DisPredict: A Predictor of Disordered Protein Using Optimized RBF Kernel.* PLoS One, 2015. **10**(10): p. e0141551.
151. Zhao, B. and B. Xue, *Decision-Tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs.* Int J Mol Sci, 2018. **19**(10).
152. Liu, Y.M., X.L. Wang, and B. Liu, *IDP-CRF: Intrinsically Disordered Protein/Region Identification Based on Conditional Random Fields.* International Journal of Molecular Sciences, 2018. **19**(9).
153. Malysiak-Mrozek, B., T. Baron, and D. Mrozek, *Spark-IDPP: high-throughput and scalable prediction of intrinsically disordered protein regions with Spark clusters on the Cloud.* Cluster Computing-the Journal of Networks Software Tools and Applications, 2019. **22**(2): p. 487-508.
154. Liu, Y., et al., *Identification of Intrinsically Disordered Proteins and Regions by Length-Dependent Predictors Based on Conditional Random Fields.* Mol Ther Nucleic Acids, 2019. **17**: p. 396-404.

155. 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.
156. 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.
157. 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.
158. Torrisi, M., G. Pollastri, and Q. Le, *Deep learning methods in protein structure prediction.* Comput Struct Biotechnol J, 2020. **18**: p. 1301-1310.
159. Pakhrin, S.C., et al., *Deep Learning-Based Advances in Protein Structure Prediction.* Int J Mol Sci, 2021. **22**(11).
160. Zhang, F., et al., *DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions.* Proteomics, 2019. **19**(12): p. e1900019.
161. Gligorijevic, V., M. Barot, and R. Bonneau, *deepNF: deep network fusion for protein function prediction.* Bioinformatics, 2018. **34**(22): p. 3873-3881.
162. Di Lena, P., K. Nagata, and P. Baldi, *Deep architectures for protein contact map prediction.* Bioinformatics, 2012. **28**(19): p. 2449-57.
163. Wang, S., et al., *Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model.* PLoS Comput Biol, 2017. **13**(1): p. e1005324.
164. Wu, T., et al., *Analysis of several key factors influencing deep learning-based inter-residue contact prediction.* Bioinformatics, 2020. **36**(4): p. 1091-1098.
165. Wang, S., S. Sun, and J. Xu, *Analysis of deep learning methods for blind protein contact prediction in CASP12.* Proteins, 2018. **86 Suppl 1**: p. 67-77.
166. Schaarschmidt, J., et al., *Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age.* Proteins, 2018. **86 Suppl 1**: p. 51-66.
167. Wu, T.Q., et al., *DeepDist: real-value inter-residue distance prediction with deep residual convolutional network.* Bmc Bioinformatics, 2021. **22**(1).
168. Sun, Z., et al., *To improve the predictions of binding residues with DNA, RNA, carbohydrate, and peptide via multi-task deep neural networks.* IEEE/ACM Trans Comput Biol Bioinform, 2021. **PP**.
169. Zhang, F., et al., *DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning.* Brief Bioinform, 2021.
170. 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.
171. Pan, X.Y. and H.B. Shen, *Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks.* Bioinformatics, 2018. **34**(20): p. 3427-3436.
172. Elbasir, A., et al., *DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction.* Bioinformatics, 2019. **35**(13): p. 2216-2225.
173. Zhang, B., L. Li, and Q. Lu, *Protein Solvent-Accessibility Prediction by a Stacked Deep Bidirectional Recurrent Neural Network.* Biomolecules, 2018. **8**(2).
174. Heffernan, R., et al., *Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.* Sci Rep, 2015. **5**: p. 11476.
175. Zhang, B., J. Li, and Q. Lu, *Prediction of 8-state protein secondary structures by a novel deep learning architecture.* BMC Bioinformatics, 2018. **19**(1): p. 293.
176. Spencer, M., J. Eickholt, and C. Jianlin, *A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction.* IEEE/ACM Trans Comput Biol Bioinform, 2015. **12**(1): p. 103-12.

177. Wang, D., et al., *MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction.* Bioinformatics, 2017. **33**(24): p. 3909-3916.
178. Li, F., et al., *DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites.* Bioinformatics, 2020. **36**(4): p. 1057-1065.
179. McGuffin, L.J., *Intrinsic disorder prediction from the analysis of multiple protein fold recognition models.* Bioinformatics, 2008. **24**(16): p. 1798-804.
180. Mizianty, M.J., et al., *In-silico prediction of disorder content using hybrid sequence representation.* BMC Bioinformatics, 2011. **12**: p. 245.
181. 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.
182. Hanson, J., et al., *Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks.* Bioinformatics, 2019. **35**(14): p. 2403-2410.
183. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.
184. Oates, M.E., et al., *D(2)P(2): database of disordered protein predictions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D508-16.
185. 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.
186. Piovesan, D., et al., *MobiDB: intrinsically disordered proteins in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D361-D367.
187. Zhao, B., et al., *DescribePROT: database of amino acid-level protein structure and function predictions.* Nucleic Acids Res, 2021. **49**(D1): p. D298-D308.
188. 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.
189. 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.
190. Meszaros, B., et al., *PhaSePro: the database of proteins driving liquid-liquid phase separation.* Nucleic Acids Res, 2020. **48**(D1): p. D360-D367.
191. 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.
192. Oldfield, C.J., Z. Peng, and L. Kurgan, *Disordered RNA-Binding Region Prediction with DisoRDPbind.* Methods Mol Biol, 2020. **2106**: p. 225-239.
193. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016.
194. 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.
195. 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.
196. 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.
197. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
198. Teufel, F., et al., *SignalP 6.0 predicts all five types of signal peptides using protein language models.* Nat Biotechnol, 2022.

199. 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.
200. Mirdita, M., M. Steinegger, and J. Soding, *MMseqs2 desktop and local web server app for fast, interactive sequence searches.* Bioinformatics, 2019. **35**(16): p. 2856-2858.
201. Steinegger, M. and J. Soding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nat Biotechnol, 2017. **35**(11): p. 1026-1028.
202. Wu, Z., et al. *Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins*. 2017. Cham: Springer International Publishing.
203. Wu, Z., et al., *Prediction of Intrinsic Disorder with Quality Assessment Using QUARTER.* Methods Mol Biol, 2020. **2165**: p. 83-101.
204. Hu, G., et al., *Quality assessment for the putative intrinsic disorder in proteins.* Bioinformatics, 2019. **35**(10): p. 1692-1700.
205. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.
206. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences.* Proteomics, 2019. **19**(6): p. e1800058.
207. 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.
208. Monzon, A.M., et al., *FLIPPER: Predicting and Characterizing Linear Interacting Peptides in the Protein Data Bank.* J Mol Biol, 2021. **433**(9): p. 166900.
209. Hanson, J., et al., *Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning.* Bioinformatics, 2020. **36**(4): p. 1107-1113.
210. Peng, Z., Q. Xing, and L. Kurgan, *APOD: accurate sequence-based predictor of disordered flexible linkers.* Bioinformatics, 2020. **36**(Supplement_2): p. i754-i761.
211. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Computational prediction of functions of intrinsically disordered regions.* Prog Mol Biol Transl Sci, 2019. **166**: p. 341-369.
212. Hosoya, Y. and J. Ohkanda, *Intrinsically Disordered Proteins as Regulators of Transient Biological Processes and as Untapped Drug Targets.* Molecules, 2021. **26**(8).
213. 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.
214. Ambadipudi, S. and M. Zweckstetter, *Targeting intrinsically disordered proteins in rational drug discovery.* Expert Opin Drug Discov, 2016. **11**(1): p. 65-77.
215. Oldfield, C.J., et al., *Utilization of protein intrinsic disorder knowledge in structural proteomics.* Biochim Biophys Acta, 2013. **1834**(2): p. 487-98.
216. Deng, X., et al., *An Overview of Practical Applications of Protein Disorder Prediction and Drive for Faster, More Accurate Predictions.* Int J Mol Sci, 2015. **16**(7): p. 15384-404.