nature protocols

Review article

https://doi.org/10.1038/s41596-023-00876-x

Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins

Received: 13 December 2022

Accepted: 21 June 2023

Published online: 22 September 2023

Check for updates

Lukasz Kurgan[1] ✉, Gang Hu[2], Kui Wang[2], Sina Ghadermarzi[1], Bi Zhao[1], Nawar Malhis[3], Gábor Erdős[4], Jörg Gsponer[3] ✉, Vladimir N. Uversky[5,6] ✉ & Zsuzsanna Dosztányi[4] ✉

Intrinsic disorder is instrumental for a wide range of protein functions, and its analysis, using computational predictions from primary structures, complements secondary and tertiary structure-based approaches. In this Tutorial, we provide an overview and comparison of 23 publicly available computational tools with complementary parameters useful for intrinsic disorder prediction, partly relying on results from the Critical Assessment of protein Intrinsic Disorder prediction experiment. We consider factors such as accuracy, runtime, availability and the need for functional insights. The selected tools are available as web servers and downloadable programs, offer state-of-the-art predictions and can be used in a high-throughput manner. We provide examples and instructions for the selected tools to illustrate practical aspects related to the submission, collection and interpretation of predictions, as well as the timing and their limitations. We highlight two predictors for intrinsically disordered proteins, flDPnn as accurate and fast and IUPred as very fast and moderately accurate, while suggesting ANCHOR2 and MoRFchibi as two of the best-performing predictors for intrinsically disordered region binding. We link these tools to additional resources, including databases of predictions and web servers that integrate multiple predictive methods. Altogether, this Tutorial provides a hands-on guide to comparatively evaluating multiple predictors, submitting and collecting their own predictions, and reading and interpreting results. It is suitable for experimentalists and computational biologists interested in accurately and conveniently identifying intrinsic disorder, facilitating the functional characterization of the rapidly growing collections of protein sequences.

For a long time, protein function was considered within the protein sequence–structure–function paradigm[1–3]. According to this paradigm, a specific function of a protein is determined by its unique three-dimensional structure encoded in its amino acid sequence. However, more recent discoveries demonstrate that many cellular functions are conducted by proteins and protein regions that do not have unique tertiary structures. These intrinsically disordered proteins and regions (IDPs and IDRs, respectively)[4–6] are relatively common

A full list of affiliations appears at the end of the paper. ✉e-mail: lkurgan@vcu.edu; gsponer@msl.ubc.ca; vuversky@usf.edu; zsuzsanna.dosztanyi@ttk.elte.hu

across all taxonomic domains[7–9]. Intrinsic disorder underlies the exceptional structural heterogeneity of proteins, which can be composed of multiple parts/segments that are folded or disordered to different degrees[10]. Disorder facilitates many key aspects of protein functions, thereby complementing functions of ordered proteins and regions[1,11–14]. Several illustrative examples of disorder-driven functions follow:

1. Enabling moonlighting (ability to carry out multiple functions)[15] and facilitating the formation of hubs in protein–protein interaction networks[16–18], including in a cell- and tissue-specific manner[19,20]
2. Contributing to cellular signaling[11,21,22] and regulation[23–26]
3. Acting as scaffolds[27–29] and being an essential part of proteinaceous machinary[30]
4. Facilitating alternative splicing and posttranslational modifications that are linked to the increased functional diversity in multicellular organisms[31,32]
5. Driving liquid–liquid phase separation and related potential to control and regulate biogenesis of various membrane-less organelles[33,34]

The interlinked and complementary presence of structured and disordered regions in protein sequences serves as a foundation for a more general protein structure–function continuum model where 'a given protein exists as a dynamic conformational ensemble containing multiple proteoforms characterized by a broad spectrum of structural features and possessing various functional potentials'[35]. This suggests that deciphering protein functions should rely on combining computational and biophysical studies of both structured regions and IDRs.

The sequence–structure–function paradigm dictates that an amino acid sequence folds into a unique structure under the physiological conditions[2]. To this end, sequences of ordered proteins are characterized by the presence of a 'folding code'[36], which can be potentially used for sequence-based prediction of protein structure. Recent advances in deep learning, including AlphaFold[37,38], confirm the utility of this code and enable accurate and high-throughput predictions of protein structures[39], with some notable exceptions[40,41]. Similarly, the lack of unique structures for IDPs/IDRs under physiological conditions is also encoded in specific features of their amino acid sequences. Early studies suggest that this 'nonfolding code' includes a low content of hydrophobic amino acids combined with elevated levels of charged residues, giving rise to the high net charges of these proteins at neutral pH[42–44]. Subsequent computational analyses of the sequences of IDPs/IDRs revealed that they are depleted in order-promoting residues, such as Trp, Tyr, Phe, Ile, Leu, Val, Cys and Asn, while being enriched in disorder-promoting Ala, Arg, Gly, Gln, Ser, Glu, Lys and Pro residues[45–47]. There are also other sequence-derived features, such as sequence complexity, that are different between the sequences of ordered proteins/domains and IDPs/IDRs[14,45,48,49]. These characteristics have fueled the development of numerous computational methods that predict intrinsic disorder from sequences[50–53].

The most comprehensive database of experimentally verified IDRs and IDPs is the DisProt database[54,55]. This database is curated from literature by a community of experts and provides information on disorder status and its functional annotations. DisProt is the key source of the ground truth for design and assessment of methods that predict disorder and specific functions of IDRs. While it is regularly updated, as of the end of 2022 it contained annotations for less than 2,500 proteins. The limited number of experimentally verified proteins highlights the importance of predictive methods for the characterization of IDPs and IDRs. This tutorial recommends and explains useful methods and resources for the prediction of disorder and disorder functions from protein sequences. We consider multiple relevant factors, including predictive accuracy, availability/convenience and runtime efficiency. We discuss availability and current location of the implementations and web servers for the selected tools. We also provide instructions for how to submit predictions and insights into how to read, interact with and interpret the results. Finally, we estimate overall runtime and runtime for specific steps of the prediction process.

Prediction of intrinsic disorder

Several surveys provide historical perspectives, categorize and describe sequence-based disorder predictors, and discuss their impact[50,52,53,56–62]. The first disorder predictor was published in 1979 (ref. 63). Since then, many different approaches have been employed for the prediction of protein disorder, ranging from amino acid scales, simplified biophysical models to more sophisticated machine learning approaches. The latest efforts focus on designing methods that rely on deep learning models, which are multilayer neural networks that typically utilize advanced network architectures (e.g., recurrent and convolutional), with an underlying objective to continue improving predictive performance[50]. The first deep learning-based disorder predictor dates back to 2013 (ref. 64), and these efforts have intensified in the last few years. More specifically, a recent review reveals that the majority of methods published since 2019 (7 out of 12) utilize deep learning models[65]. Several large-scale studies were carried out to comparatively evaluate predictive quality of intrinsic disorder predictors and to assess progress[66–76]. These studies include community-driven assessments where predictive methods are evaluated on blind test datasets (i.e., proteins that were not available to the authors of predictors) by assessors who do not take part in the competitions. These assessments include Critical Assessment of Structure Prediction (CASP) between CASP5 to CASP10 (refs. 71–76) and Critical Assessment of Intrinsic protein Disorder (CAID) that was published in 2021 (ref. 70). The CAID experiment was the largest to date, involved 32 methods, and evaluated their predictive accuracy and runtime. It found that the best performing methods[70,77], which include flDPnn[78], SPOT-Disorder2 (ref. 79), RawMSA[80] and AUCpred[81], rely exclusively on deep neural networks. Interestingly, these methods utilize a variety of neural network architectures, such as feed-forward (flDPnn), convolutional (AUCpred), recurrent (SPOT-Disorder2) and a hybrid of convolutional and recurrent (RawMSA). A recent study echoes these results and empirically demonstrates that the deep learning-based predictors statistically outperform other types of predictors[65], which partly explains the focus on this predictive model.

The deep learning-based method that recently shook the protein structure prediction field, AlphaFold2 (ref. 37), also provides results that can be used to identify intrinsic disorder. For instance, low values of the predicted local distance difference test scores, which estimate reliability of the AlphaFold2's structure predictions at the residue level, and window-based averaging of the predicted relative solvent accessible surface have been shown to predict IDRs relatively accurately on the DisProt datasets[82,83]. Moreover, AlphaFold2 was combined with Rosetta ResidueDisorder[84] to produce a disorder prediction[85]. However, recent studies show that these results take much more time to produce and are not as accurate as the predictions produced by other disorder predictors[86–88].

We compiled results generated on the DisProt dataset from the CAID experiment[70] and a subsequent study that assesses newer methods on the same dataset[65] to analyze predictive performance in the context of when these methods were released, their predictive models and runtime. We focus on methods that are publicly available as either standalone code and/or a web server, which ensures that the end users can relatively easily collect their results. We use the recently released large benchmark dataset from the CAID experiment that includes 646 proteins and excludes proteins with ambiguous disorder annotations. The results in CAID were measured using bootstrapping with 1,000 repetitions and the assessment was blind (i.e., disorder annotations were not publicly available at the time) and so predictors could not be trained on these data. The runtime was evaluated on the same equipment that includes Intel 8 core processors with 16 GB of random access memory and the Ubuntu 16.04 operating system. We quantify predictive

https://doi.org/10.1038/s41596-023-00876-x

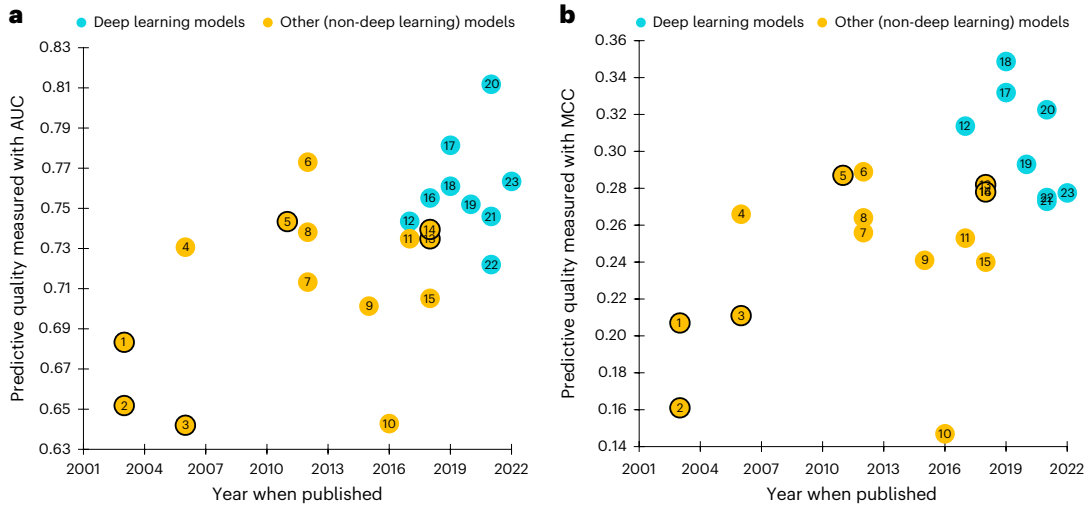

Fig. 1 | Predictive performance of disorder predictors available to end users. The results are computed based on the benchmark dataset from CAID[70] that was also used in a subsequent study[65]. The methods are divided into two groups: those that rely on deep learning models (blue markers) versus those that utilize other types of models (orange markers). **a**, The AUC values. **b**, The MCC values. Methods identified using markers with black border are fast, i.e., on average they complete prediction for a single protein in under 1 s. The predictors are encoded as follows: (1) DisEMBL-465 (ref. 148), (2) DisEMBL-HL[148], (3) FoldUnfold[181], (4) PONDR VSL2B[142], (5) IsUnstruct[182], (6) Espritz-DisProt[147], (7) Espritz-NMR[147], (8) Espritz-XRay[147], (9) DISOPRED3 (ref. 183), (10) DisPredict[184], (11) MobiDB-lite[122], (12) SPOT-Disorder[165], (13) IUpred-long[112] (v2), (14) IUpred-short[112] (v2), (15) pyHCA, (16) SPOT-Disorder-Single[168], (17) RawMSA[80], (18) SPOT-Disorder2 (ref. 79), (19) IDP-Seq2Seq[166], (20) flDPnn[78], (21) Metapredict[167], (22) RFPR-IDP[169] and (23) DisoMine[185]. Additional details for these predictions are in Supplementary Table 1.

performance with two popular metrics: the area under receiver operating characteristic curve (AUC) and Matthew's correlation coefficient (MCC), which we define in Supplementary Information.

Figure 1 summarizes the results for 23 publicly available disorder predictors that were tested on the CAID dataset[65,70]. We find a clear upward trend in predictive performance relative to the publication time. This confirms observations from recent studies that point to a steady progress in improving predictive quality as newer tools are released[50,89]. We also observe that deep learning-based tools produce more accurate predictions, which agrees with recent studies[65,70]. As it was observed in CAID[70], some of the more recent methods produce relatively accurate results, with AUC values >0.75 and MCC values >0.3. Figure 1 identifies several accurate methods, including flDPnn, RawMSA and SPOT-Disorder2. Runtime data published in CAID reveal that they on average take about 20 s, 250 s and 2,000 s to predict one protein, respectively[70]. Moreover, RawMSA is available only as source code, while the other two tools have both source code and web server options. Altogether, when considering predictive performance, runtime and availability, the currently best option is flDPnn.

We separately consider very fast predictors. This aspect is particularly relevant for applications where disorder predictions are generated for large collections of proteins. Numerous examples of such large-scale studies are available, including recent analysis of RNA-binding proteins in a human proteome[90], investigation of distribution of intrinsic disorder across subcellular compartments[91], analysis of coronaviruses[92,93] and other viral proteomes[94], and identification of cancer driver genes[95,96]. Five methods are capable of predicting a given protein in under 1 s: DisEMBL-465, DisEMBL-HL, FoldUnfold, IsUnstruct, IUPred-long and IUPred-short. The first three tools offer relatively low levels of predictive performance (Fig. 1), and IsUnstruct is available as only a web server. Consequently, the best option to quickly generate disorder predictions is IUPred.

The two methods that we recommend provide a good trade-off between predictive performance and runtime, with flDPnn being about one order of magnitude slower but producing more accurate predictions, while IUPred is extremely fast and reasonably accurate. Moreover, both tools are available as source code that can be installed and run on user's computers, and as user-friendly web servers. In the latter case, predictions are done on the server side and the results are returned to the user in two complementary formats: a text file that can be downloaded and parsed to collect the predictions and a graphical file that visualizes the predictions.

Prediction of disordered binding regions

One of the key functional mechanisms of IDPs is molecular recognition, which involves a variety of binding modes. Some IDRs contain molecular recognition features (MoRFs), short regions capable of binding-induced folding that are implicated in signaling and regulation[97–101]. Plasticity of IDRs also enables them to fold differently when interacting with different partners[102]. This can give rise to binding promiscuity in the form of one-to-many or many-to-one interactions[102,103]. Some IDRs form dynamic complexes[104], including 'fuzzy' assemblies with high levels of disorder in the bound state[105–107]. Moreover, individual binding IDRs may overlap and form molecular switches[108]. Interestingly, binding IDRs are also characterized by unique biases of their amino acid sequences[47], making them predictable from sequence.

While there are over 100 disorder predictors available and many of them continue to be frequently used[50], recent research has shifted to building methods that predict specific functional types of IDR, in particular those that interact with ligands. A few surveys summarize recently developed predictors of disordered binding regions[53,100,109–111]. With over three dozen of these methods[111], the majority of them focus on predicting MoRFs[97–101]. Several methods also target prediction of a more generic class of disordered protein-binding regions, which are not limited to shorter segments[100]. The impact and value of these methods are reflected by the inclusion of their evaluation in the CAID experiment[70]. Disordered binding regions were defined according to the DisProt database as regions that were shown to be involved in binding based on experiments. CAID evaluated 11 predictors of disordered binding regions and found that five of them perform above a baseline level: ANCHOR2 (ref. 112), DisoRDPbind[113], MoRFchibi$_{Light}$ (ref. 114), MoRFchibi$_{Web}$ (ref. 115) and OPAL[116]. The best method that targets prediction of disordered protein-binding regions is ANCHOR2; it secures an AUC of 0.742 and MCC of 0.199 in CAID[70]. The highest scoring

https://doi.org/10.1038/s41596-023-00876-x

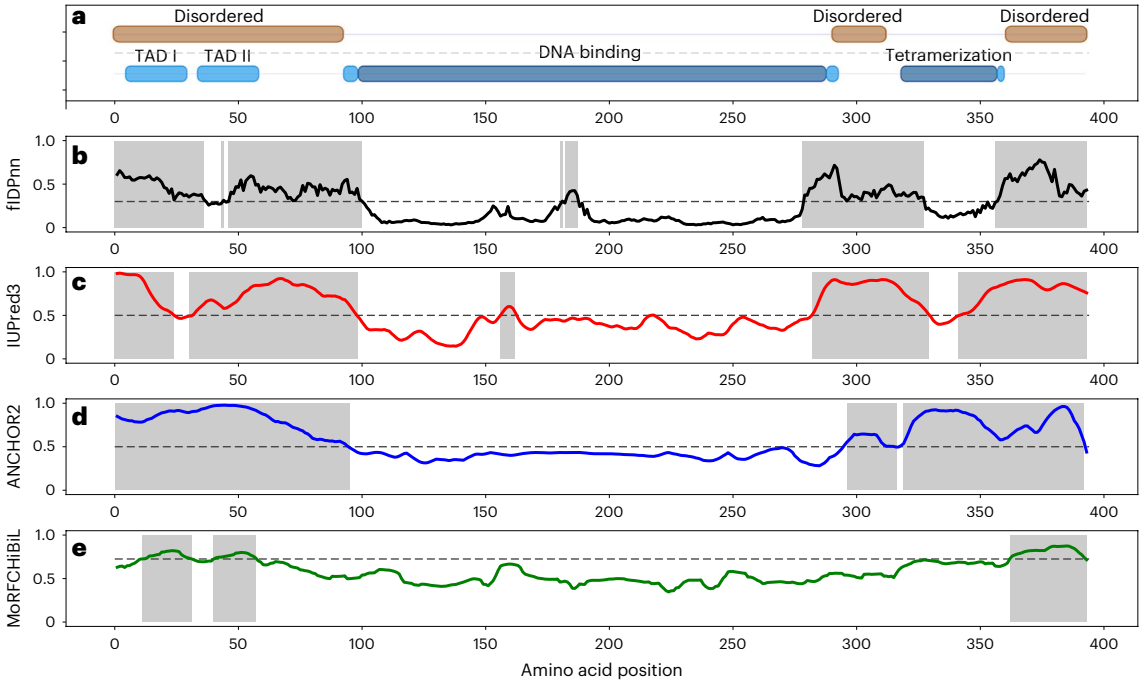

Fig. 2 | Computational analysis of human cellular tumor antigen p53 (UniProt[158] ID: P04637) using disorder and disorder function predictors. **a**, Color-coded native annotations of disorder (brown) and domains (blue) collected from the DisProt database. For the domain annotations: light blue denotes annotations sourced from either Pfam or Gene3D versus dark blue for annotations where both sources overlap. **b**, Disorder prediction by flDPnn[78]. **c**, Putative disorder profile generated by IUPred (v3)[121]. **d**, Prediction of binding IDRs generated by ANCHOR2. **e**, MoRF prediction profile generated by MoRFchibi$_{Light}$ (ref. 114).

predictor of MoRFs is MoRFchibi$_{Light}$, which obtains an AUC of 0.720 and MCC of 0.161 in CAID[70]. Both methods are fast and conveniently available to the end users as source code that can be installed locally and web servers that can be used remotely. CAID also reports that ANCHOR2 is very fast with runtime <1 s per protein while MoRFchibi$_{Light}$ takes on average ~3 s per protein[70]. Given their favorable predictive performance, availability and short runtime, we recommend ANCHOR2 and MoRFchibi$_{Light}$ for the prediction of disorder functions. In general, ANCHOR2 predicts binding regions that are longer while MoRFchibi$_{Light}$ predicts shorter MoRF regions.

Examples

We first demonstrate the usefulness of computational predictors of disorder and disordered binding regions for the human cellular tumor antigen p53, which aptly illustrates the structure–function continuum model[117]. This protein has multiple domains (Fig. 2a): the intrinsically disordered N-terminal region that hosts translational activation domains (residues 1–93), the structured central DNA-binding domain (100–288), a short IDR (291–312) followed by the tetramerization domain (319–357) and disordered C-terminal region and regulatory domain (359–393). We collect these domain and disorder annotations from the DisProt database[54], where the domain annotations are extracted from Pfam[118] and Gene3D[119]. Note that Pfam annotations correspond to sequence families that are not necessarily ordered. Furthermore, the tetramerization domain can also undergo an order-to-disorder transition and expose nuclear export signal[120]. Figure 2 shows predictions of disorder and binding regions generated solely from the p53 sequence by popular and relatively accurate methods that we selected for this tutorial, flDPnn[78], IUPred[121], ANCHOR2 (ref. 112) and MoRF$_{CHiBi}$ (ref. 114). Each prediction consists of numeric propensities and binary scores for each amino acid from the input protein sequence. Higher values of propensities indicate higher likelihood that the corresponding amino acid is intrinsically disordered or belongs to a disordered binding region. The binary scores categorize each amino acid as either disordered versus ordered (for disorder predictions from IUPred and flDPnn) or binding versus nonbinding (for predictions from ANCHOR2 and MoRF$_{CHiBi}$). Binary predictions are typically produced from the propensities using a threshold, i.e., residues with propensities exceeding threshold are marked as disordered/disordered binding.

Results from flDPnn include the black line for propensities, horizontal dotted line for the threshold of 0.3 and gray-shaded vertical bands for the binary prediction of disorder, where the propensity exceeds the threshold (Fig. 2b). They suggest presence of an IDR at the N-terminus (residues 1–100), a short IDR in the middle (181–187) and two IDRs close to the C-terminus (279–327 and 357–393). Predictions from IUPred[121] are composed of the red line for propensities, threshold line at 0.5 and the gray-shaded bands for the binary prediction of disorder (Fig. 2c). Similar to flDPnn, the IUPred's predictions imply presence of IDRs at both termini and an ordered region in the middle. We note that both disorder predictions are in good agreement with each other and with the native annotations, closely overlapping with the experimentally confirmed IDRs (Fig. 2a). However, some predictions are incorrect. For instance, flDPnn misses a fragment of the N-terminus IDRs near position 40 and incorrectly predicts a short IDR near position 185. These predictions are associated with the putative propensities near the threshold value (dotted horizontal line in Fig. 2b). We note a similar pattern for the IUPred's predictions, where the incorrect predictions near positions 30, 160 and 220 have propensities that are also close to the threshold. More generally, disorder predictions associated with high and low values of propensities are more likely to correctly identify disordered and structured regions, respectively, while predictions with propensity scores near the threshold should be considered as less certain. Another way to increase confidence in a given prediction is to cross-check it against another prediction. For instance, sequence regions where predictions of flDPnn and IUPred agree, i.e., both predict disorder or both predict order, tend to identify

https://doi.org/10.1038/s41596-023-00876-x

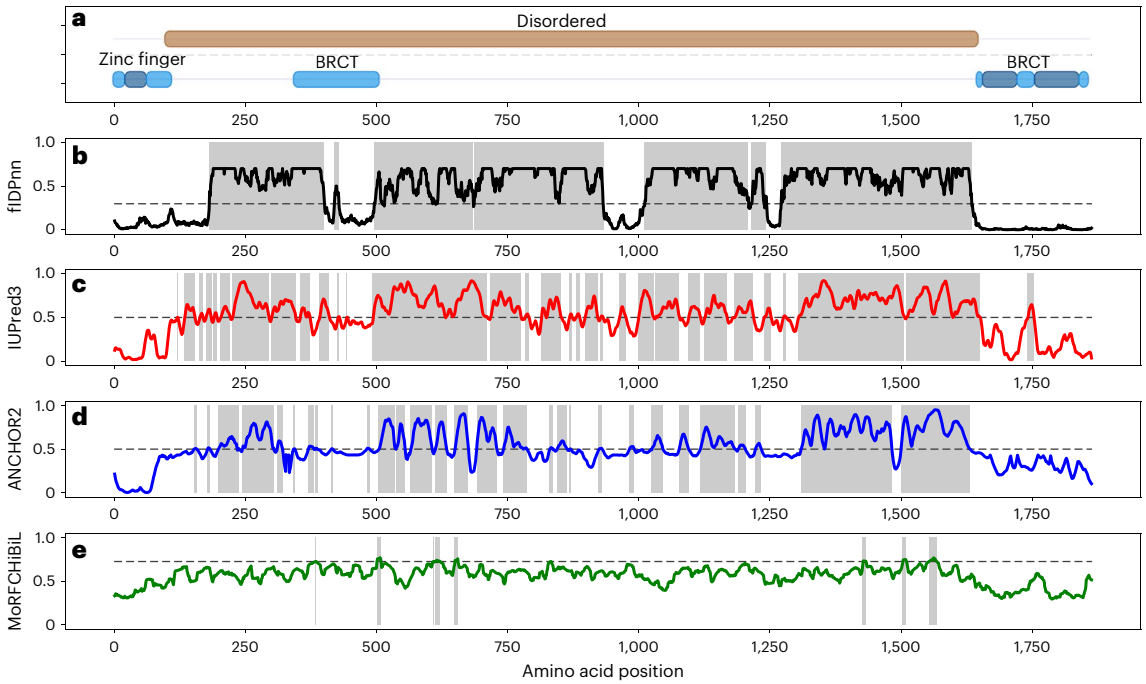

Fig. 3 | Computational analysis of human BRCA1 (UniProt[158] ID: P38398) using disorder and disorder function predictors. a, Color-coded native annotations of disorder (brown) and domains (blue) collected from the DisProt database. For the domain annotations: light blue denotes annotations sourced from either Pfam or Gene3D versus dark blue for annotations where both sources overlap. **b**, Disorder prediction by flDPnn[78]. **c**, Putative disorder profile generated by IUPred (v3)[121]. **d**, Prediction of binding IDRs generated by ANCHOR2. **e**, MoRF prediction profile generated by MoRFchibi$_{Light}$ (ref. 114).

regions that are predicted more accurately. This underlies the design of tools that implement a consensus of results from multiple disorder predictors, which are shown to be on average better when compared with the corresponding results generated by the corresponding individual predictors[122–124].

At the disordered transactivation regions at the N-terminus, p53 interacts with TFIID, TFIIH, Mdm2, RPA, CBP/p300 and CSN5/Jab1 among many other proteins[125], whereas its C-terminal domain acts as a binding hub for GSK3β, PARP-1, TAF1, TRRAP, hGcn5, TAF, 14-3-3, S100B(ββ) and many other proteins[125]. These lists of p53 interactors represents a small subset of almost 1,000 known partners of this protein[117]. In effect, the corresponding binding regions overlap and cover a large portion of the native IDRs. In agreement with these annotations, ANCHOR2's prediction (Fig. 2d; the blue line for propensities, threshold line at 0.5 and the gray-shaded bands for the binary prediction of protein-binding IDRs), accurately suggests that IDRs interact with proteins. Predictions from MoRF$_{CHiBi}$ (Fig. 2e; the green line for propensities, threshold line at 0.725 and the gray-shaded bands for the binary predictions) identify three putative MoRF regions: residues 11–31, 40–57 and 362–392. Importantly, these regions coincide with validated binding sites and functional motifs of p53. For example, the disordered transactivation domains I and II (TAD I and TAD II motifs) are located at regions 6–30 and 35–59, respectively (Fig. 2a).

The second example considers the breast cancer type 1 susceptibility protein (BRCA1), which is much longer than p53 (1,863 versus 393 amino acids), has a long IDR in the middle and no IDRs at the termini (Fig. 3). This protein has only two relatively small structured domains, the N-terminal zinc finger RING (Really Interesting New Gene) domain (residues 1–109) and two C-terminally located tandem copies of the BRCA1 C-terminal domain (BRCT1 and BRCT2, residues 1,642–1,736 and 1,756–1,855, respectively)[126]. The long IDR (residues 100–1,649) contains the serine-rich domain associated with BRCT (residues 345–508), two nuclear localization sequences (NLS, 503–508 and 607–614), a serine cluster domain (1,280–1,524) and a coiled-coil domain (1,367–1,437)[127]. Figure 3 shows that in agreement with the aforementioned experimental data, the central BRCA1 region is predicted to have very high levels of intrinsic disorder. For example, as per flDPnn, most of the residues within the 182–400, 496–932, 1,012–1,243 and 1,272–1,634 regions are predicted as disordered (Fig. 3b). IUPred mostly agrees with these observations and shows high disorder content for the region that coincides with the experimentally validated IDR (Fig. 3c).

Importantly, this central region of BRCA1 was experimentally shown to act as an intrinsically disordered scaffold for multiple protein–protein and protein–DNA interactions[128]. In fact, as per DisProt annotations, human BRCA1 was shown to possess several disorder-based protein binding regions, such as residues 175–394, 433–511, 740–1,083 and 1,343–1,440 (DisProt ID: DP00238). Furthermore, other subregions of this central IDR were shown to interact with various proteins; e.g., 341–748 interacts with the growth arrest and DNA damage-inducible protein GADD45 alpha and DNA damage repair protein RAD50 (refs. 127,129), whereas the C-terminal part of this central region together with the BRCT domains are engaged in interactions with BRCA2 (1,314–1,863), histone deacetylase complex (HDAC1 and HDAC2, 1,536–1,863), RNA helicase A (1,560–1,863) and CtBP-interacting protein (1,561–1,863)[127,129]. Figure 3d shows that these experimentally validated binding regions align with the predictions from ANCHOR2. For example, 175–394 and 740–1,083 regions are predicted to contain eight putative binding regions each, with the longest regions being 199–234, 246–305, 743–786 and 1,026–1,046. Moreover, the entire experimentally validated region 1,343–1,440 is predicted as protein binding (binary prediction at 1,310–1,482 in Fig. 3d). Similarly, Fig. 3e shows that, as per MoRFchibi$_{Light}$, the central region of BRCA1 contains eight MoRFs (residues 385, 502–508, 609, 613–620, 650–655, 1,428–1,433, 1,504–1,510 and 1,554–1,567). Three of those MoRFs (502–508, 609 and 613–620) coincide or overlap with the two known NLSs of BRCA1 (503–508 and 607–614). These data, taken together, indicate that ANCHOR2 and MoRFchibi$_{Light}$ are capable of predicting disorder-based protein binding sites and NLS motifs in

BRCA1. As expected, $MoRF_{CHiBi}$ focuses on finding shorter binding regions that fold upon binding (MoRFs) while ANCHOR2 predicts longer IDRs involved in binding.

Selected methods

In the following, we provide details on the methodology, availability, details of the web server access, and practical aspects related to limitations and use of the selected four methods.

Accurate prediction of intrinsic disorder with flDPnn

We recommend flDPnn[78] for projects that analyze individual proteins or relatively small protein sets. This method relies on a comprehensive sequence-derived input feature space and a relatively simple deep fully connected feed-forward neural network. The major types of inputs include initial disorder prediction generated by IUPred (v1)[130] that is refined and improved by flDPnn, disorder function predictions produced by DisoRDPbind[113,131], DFLpred[132] and fMoRFpred[99], putative secondary structure produced by single-sequence version of PSIPRED[133], and evolutionary information generated by PSI-BLAST[134] using a small Swiss-Prot database. These inputs provide a broad coverage of information relevant to disorder prediction and are generated quickly from the sequence (e.g., PSI-BLAST uses a small database). The two innovations that underly flDPnn are the use of predicted disorder functions and the formulation of the protein-level features that quantify bias of the whole protein to be disordered[78]. Ablation analysis shows that versions of the flDPnn model that exclude one of the typical inputs (e.g., model without evolutionary input or without disorder prediction from IUPred) produce similarly high levels of predictive performance as the original flDPnn, while exclusion of the novel inputs results in a more substantial drop in performance[78].

There are three options to use this tool, all available for free for academic use:

- Web server version at http://biomine.cs.vcu.edu/servers/flDPnn/. This version makes predictions on the online server side, without the need to install any software, and is arguably the most convenient to use. Box 1 and Fig. 4 describe details how to perform predictions and analyze the corresponding results when using the web server
- Source code available at https://gitlab.com/sina.ghadermarzi/fldpnn. The code can be run on a user's computer system with Linux operating system (Ubuntu x64 20.04.2 or newer preferred) with tcsh shell (6.21.00-1 or newer), Java Runtime Environment (openjdk 1.0.8 or newer) and Python 3 (3.8.5 or newer) that includes the following packages: plotly (4.14.3 or newer), scikit-learn (0.23.1 or newer), keras (2.4.3 or newer), tensorflow (2.4.1 or newer) and pandas (1.2.2 or newer). As described in a README.md file, users need to download the code, unzip it and run the following command: *python3 run_fldpnn.py protein.fasta* where protein.fasta is a FASTA-formatted file that contains sequences of input proteins. The output consists of two files: results.csv and results.html, which are in the same format as the results generated by the web server (Box 1). These files should be copied or renamed because the subsequent predictions overwrite them. In contrast to the web server, there is no limit to the number of sequences in the input file
- Docker version of the source code available at https://gitlab.com/sina.ghadermarzi/fldpnn_docker. This option also runs in the Linux-based environments, requires the use of the docker application and it is arguably easier to install compared with the source code option

High-throughput prediction of intrinsic disorder with IUPred

We recommend the IUPred predictor[121] to secure quick disorder prediction or to analyze large protein sets. The newest version 3 of the popular IUPred algorithms[121] offers improved visualization options and additional smoothing function compared with the previous version[112] that was evaluated in CAID; the predictive performance was shown to be slightly better for the newest version[121]. IUPred estimates whether a residue in the input sequence is able to form favorable interaction with its local environment, where this estimate serves as a proxy for

BOX 1

Prediction of disorder with the flDPnn web server

Submission of query sequence(s)

● **TIMING 2 min**

Navigate to the http://biomine.cs.vcu.edu/servers/flDPnn/ website (Fig. 4a). Provide the query amino acid sequences in the FASTA format (label 1 in Fig. 4a). Up to 20 sequences can be submitted for a single prediction job. Optionally, input an email address in the text box to receive email with a link to the results when the job is completed (label 2 in Fig. 4a). Click 'Run flDPnn' to submit the prediction job.

Job monitoring

● **TIMING 30 s to 1 h**

Once the job is submitted, it is put into a queue of jobs submitted to the biomine.cs.vcu.edu server and the browser redirects to a page that displays a confirmation, a job ID number (label 1 in Fig. 4b) and a location where the results will be produced (label 2 in Fig. 4b). The server processes several jobs in parallel and limits their sizes so they do not block access to other users. Thus, since flDPnn's prediction takes on average 15 s per sequence, the number of input sequences is limited to 20. When the prediction is completed, an email notification with the link to the results page is sent and the browser window is redirected to the page that provides links to the results (Fig. 4c). The results are available in two formats: (1) an html page that provides graphical view of the results (label 1 in Fig. 4c) and (2) a text file that provides raw predictions (label 2 in Fig. 4c).

Accessing and reading results

● **TIMING 3 min per submitted protein sequence**

The text-formatted file (linked at label 1 in Fig. 4d) includes a header with formatting instructions followed by the raw predictions in the comma-separable format. This file can be used to parse and retrieve the raw data for further processing and use.

The results page has interactive panels that provide graphical view of the predictions for each submitted protein (labels 2 and 3 in Fig. 4d). These panels show putative propensities for disorder using the black lines at the top, the binary disorder prediction using black horizontal bars and propensities to bind proteins, bind DNA, bind RNA and be linkers for the predicted intrinsically disordered residues that are represented using color-coded horizontal bars. The binding and linker propensities are encoded such that higher propensity values are denoted by darker shades of colors. The panels are interactive and allow for zooming in on a specific fragment of a given sequence by holding the left mouse button (double left-button click restores full sequence view) and reading the predicted propensity values and amino acids in the sequence by hovering over the corresponding lines and bars. Users can turn on and off each of the predictions by clicking on the corresponding entries in the legend (label 4 in Fig. 4d). They can also take a screenshot of the graphical panel, as well as pan, auto-scale and reset the axes of the panel using the menu in the top-right corner (label 5 in Fig. 4d).

https://doi.org/10.1038/s41596-023-00876-x

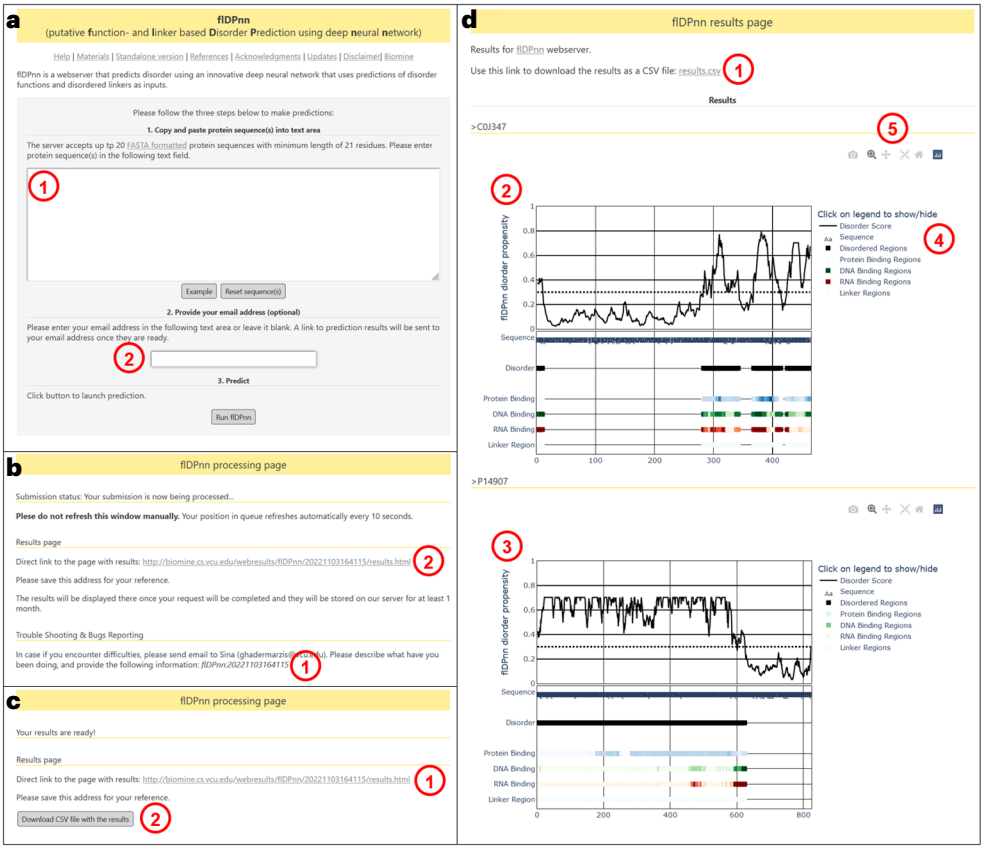

Fig. 4 | Example of flDPnn web server pages. a, The interface/input page where users submit their protein sequences. **b**, The processing page that indicates the status of the submitted prediction job. **c**, The ready page that indicates that prediction was completed and provides location of the results. **d**, The graphical page with examples results/outputs.

disorder prediction. The energy estimation is achieved by a statistical potential-based force field. Residues that exhibit favorable estimated energies tend to reside in ordered regions, whereas amino acids lacking sufficient interaction energies are prone to be intrinsically disordered. This relatively simple approach effectively captures basic biophysical principles underlying order and disorder in proteins while being fast to compute, allowing for predicting whole proteomes in minutes on a desktop computer. Users can choose different prediction modes. The default option is 'IUPred long disorder', which focuses on the identification of longer IDRs with likely biological relevance. Additional options focus on identifying disorder defined by missing residues in Protein Data Bank (PDB) structures ('IUPred short disorder'), finding structured domains ('IUPred structural domains') or predicting regions that might undergo a disorder to order transition upon the change in environmental redox potential ('Redox state'), Fig. 5a. Users can also choose between different smoothing options. Currently there are three ways to use IUPred:

- Web server at https://iupred.elte.hu/, which generates predictions on the server side, without the necessity to install any software, and sends the results to the user. Box 2 provides information about generating and analyzing predictions using the web server
- RESTFUL API version, which allows for the programmatic use of the web server
- Standalone package downloadable from https://iupred.elte.hu/download_new. The package is free for academic use. IUPred requires the Python3 interpreter and the 'scipy' package[135] to be installed. The software contains an executable Python script that allows users to analyze the FASTA-formatted sequences, as well as an importable Python3 library, which facilitates integration of IUPred into bioinformatics workflows

An alternative to making the disorder predictions is to collect precomputed predictions from one of three available databases: the Database of Disorder Protein Predictions (D^2P^2)[136] at https://d2p2.pro/, MobiDB[137–140] at https://mobidb.bio.unipd.it/ and DatabasE of StruCtuRe and functIon residue-Based prEdictions of PROTeins (DescribePROT)[141] at http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/. These resources cover large collections of proteins ranging from 2.26 million proteins from 273 proteomes in DescribePROT, 10.43 million proteins from 1,765 proteomes in D^2P^2, to 219.74 million proteins in MobiDB. Their key advantage is the ability to instantaneously retrieve already precomputed and stored predictions. They also reduce wasteful replication of predictions where the same method is tasked to make predictions for the same protein submitted by different users. DescribePROT provides predictions generated by an older PONDR VSL2B predictor[142]. D^2P^2 delivers predictions from nine, also mostly older methods that were published in 2012 or earlier: PONDR VL-XT[143], IUPred-short[144], IUPred-long[144], PONDR VSL2B[142], PrDOS[145], PV2 (ref. 146), ESpritz-NMR[147], ESpritz-Xray[147] and ESpritz-DisProt[147]. MobiDB similarly relies on nine tools published in 2012 or earlier: DisEMBL-HL[148], DisEMBL-465 (ref. 148), GlobPlot[149], IUPred-short[144], IUPred-long[144], PONDR VSL2B[142], ESpritz-NMR[147], ESpritz-Xray[147] and ESpritz-DisProt[147]. Given that MobiDB and D^2P^2 store multiple disorder predictions, they also produce a consensus prediction to provide a single, ultimate result. MobiDB computes the consensus using the MobiDB-lite algorithm[122] and D^2P^2 applies a 75% consensus approach, i.e., an amino acid is predicted as disordered if at least 75% of methods predict it as disordered. We note that D^2P^2 has not been updated since 2015 and is no longer actively maintained. DescribePROT covers a relatively small number of proteins since its focus is to provide access to multiple and diverse types of predictions, which besides intrinsic disorder include binding

https://doi.org/10.1038/s41596-023-00876-x

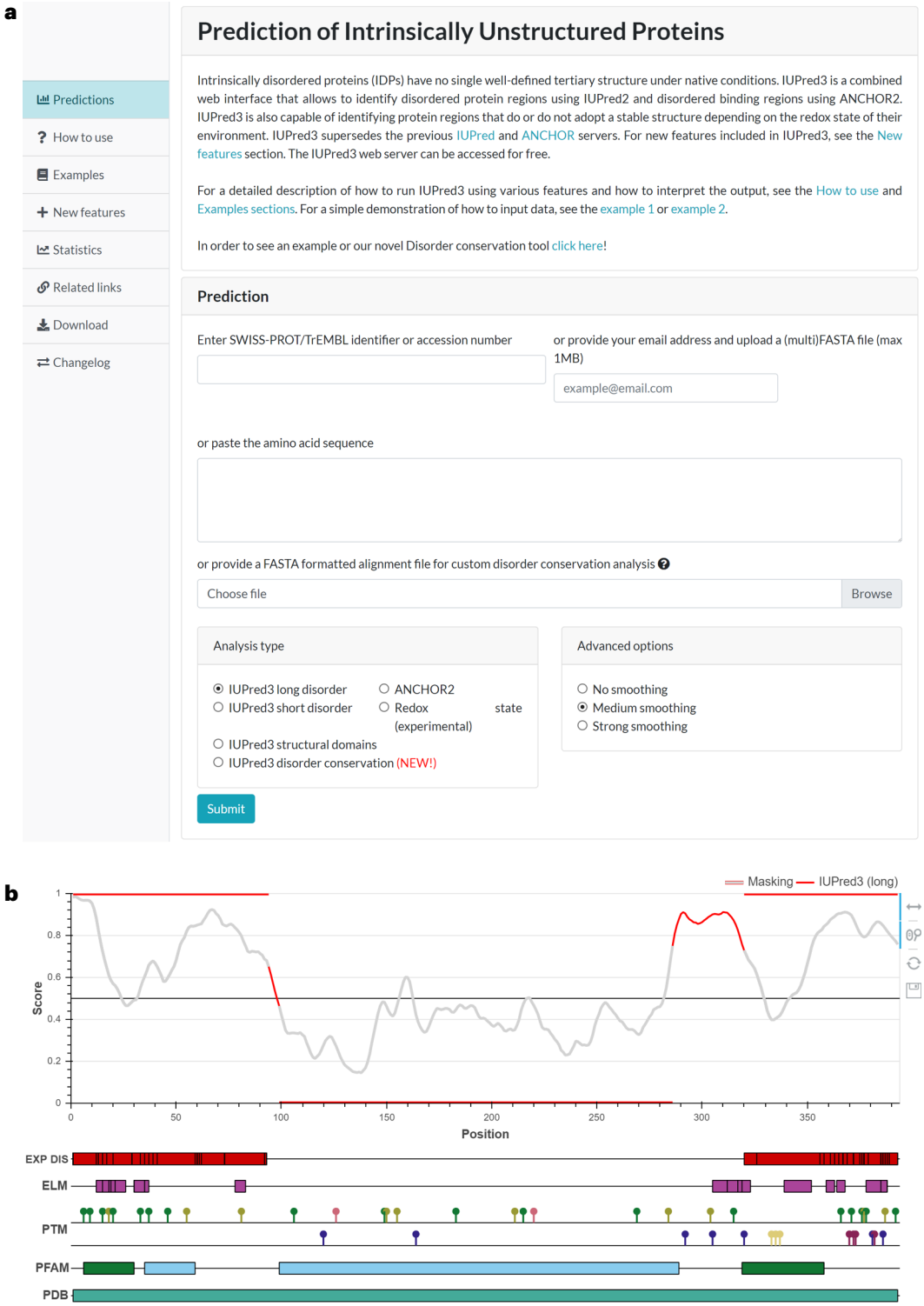

Fig. 5 | Example of IUPred web server pages. a, Home page interface of the web server. **b**, Graphical representation of the predictions for the human p53 protein (UniProt ID: P04637) with additional annotations. Regions with experimental status are highlighted by red line. This can be toggled using the 'Masking' button on the top right corner.

residues, secondary structure, signal peptides, linkers, solvent accessibility and sequence conservation. More detailed comparisons of these resources are available in a recent survey article[61].

Out of the three choices, we recommend the largest MobiDB database. MobiDB facilitates collection of predictions for individual proteins, which are provided in several parsable text formats (json, tsv, fasta) and in an interactive graphical format, as well as for user-defined datasets of proteins. The dataset can be extracted in a variety of ways, including selecting whole proteomes, sequences of given length range, preclustered protein sets and others. MobiDB is cross-linked and includes experimental data from ten external sources: CoDNaS[150], DIBS[120], DisProt[151], ELM[152], FuzDB[153], IDEAL[154], MFIB[155], PDBe[156], PhasePro[157] and UniProt[158]. These data provide useful functional and structural context for the disorder predictions. One of the

https://doi.org/10.1038/s41596-023-00876-x

BOX 2

Prediction of disorder and disorder functions with the IUPred/ANCHOR2 web server

Submission of query sequence(s)

● **TIMING 2 min**

Navigate to the website of the current version of IUPred at https://iupred.elte.hu (Fig. 5a). This is a rolling web domain, meaning that this address provides access to the latest version of IUPred (currently v3). To analyze a single protein, use either the 'Enter SWISS-PROT/TrEMBL identifier or accession number' input box with a UniProt[158] accession, or provide an amino acid sequence in the 'or paste the amino acid sequence' box. The sequence can be formatted as either plain text or in the FASTA format. To analyze multiple proteins, upload a multi-FASTA formatted file using the 'provide a FASTA formatted alignment file for custom disorder conservation analysis' input box. If this option is selected, an email address must be supplied in the respective box. Select a specific type of the requested analysis using the 'Analysis type' box, or optionally select a prediction of a conditional/functional disorder type. Users can also select from smoothing options for the disorder prediction. 'No smoothing' is equivalent to IUPred v2. Click the 'Submit' button to start the prediction.

Job monitoring

● **TIMING 5 s**

Job monitoring step is not needed because IUPred is a very fast method.

Accessing and reading results

● **TIMING 3 min per submitted protein sequence**

The graphical representation of the results consists of two main parts (Fig. 5b): an interactive line plot of the disorder prediction at the top of the page and a panel at the bottom of the page with additional information for the submitted protein. This additional information can guide the interpretation of the prediction. The first line (EXP DIS, red horizontal bar) provides experimental disorder data derived from the DisProt database[54]. The second line (ELM, purple horizontal bar) gives known linear motifs collected from the ELM database[186]. The third and fourth lines (PTM, color-coded lollipop representation) show posttranslational modification sites from the PhosphoSitePlus database[187]. The fifth line (PFAM, color-coded horizontal boxes) provides annotations from the PFAM database[188], where colors identify different types of annotation (domains, families, repeats, etc.). The last line (PDB, green horizontal bar) displays combined coverage by available PDB structures[189]. The 'Show structures' checkbox expands this section to show structures individually.

If the 'IUPred structural domains' option was selected, then a text box will appear below the graphical interface. This box includes the number of predicted globular domains and their locations alongside a string representation of the protein sequence, where capital letters represent structured domains.

If the 'Redox state (experimental)' option was selected, then a different prediction plot will be shown. This plot contains two lines (redox plus and redox minus) that correspond to the oxidative and reductive environments. Regions predicted for the redox sensitive conditional disorder are marked with a red background.

The selected prediction can be downloaded in the text and json formats.

key limitations of MobiDB and the other two databases is that they are limited to protein sequences that are included in these resources. This means that users must rely on disorder predictors for novel sequences and proteins that are not yet included in a given database. Moreover, MobiDB utilizes MobiDB-lite's predictions, which offer predictive accuracy that is similar to the results from IUPred (Fig. 1), and provides only

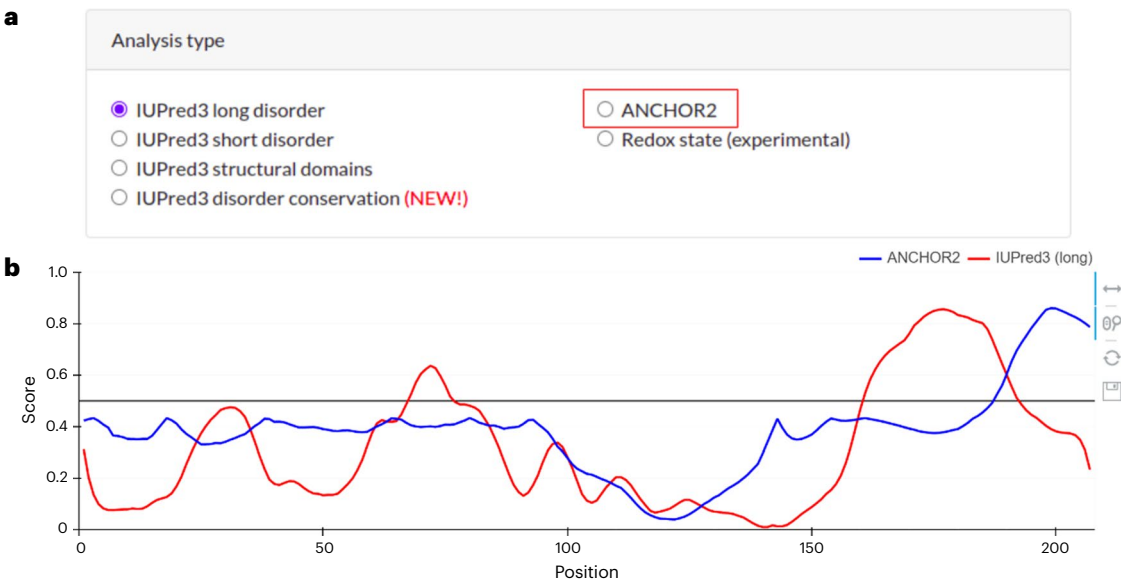

Fig. 6 | Example of ANCHOR2 web server pages. a, Selection of ANCHOR2 in the 'Analysis type' window. **b**, Graphical representation of the predictions for the human T-cell surface glycoprotein CD3 epsilon chain (UniProt ID: P07766). The blue line shows the disordered binding region prediction by ANCHOR2, the red line is the disorder prediction output by IUPred (v3). This protein contains binding site at the C-terminal region which is correctly captured by ANCHOR2's predictions.

https://doi.org/10.1038/s41596-023-00876-x

BOX 3

Prediction of disorder functions with the MoRFchibi SYSTEM

Submission of query sequence(s)

● **TIMING 2 min**

Navigate to the MoRFchibi SYSTEM website at https://morf.msl.ubc.ca/index.xhtml (Fig. 7a) and provide the query amino acid sequences in the FASTA format. Click 'Submit job'. The optional specification of the user email address allows the server, once the prediction is completed, to send a 'results ready' email with a link to the results page, which makes it safe to disconnect after submitting the job. MoRFchibi SYSTEM utilizes a separate private queue for each user. Only two sequences per user are inserted into the server queue at a time. This two-tier queue[114] adopted by the MoRFchibi server prevents one or a few jobs from dominating the server, and thus no explicit limit is placed on the number of input sequences.

Job monitoring

● **TIMING 20 s to 40 min**

Once query sequences are submitted, they appear in the jobs table. Each line in the jobs table represents a single query sequence and includes seven fields: (1) id: job ID that is system-generated; (2) label: the FASTA header; (3) size: the number of amino acids in the query sequence; (4) submitted: the date and time the query sequence entered the server queue; (5) status: job status that can have one of the following values (Fig. 7a): pending (query is in the user's private queue), position *x* (query is at position *x* in the server queue), processing (query sequence is being processed), completed in *x* s (time used to complete the prediction), error (errors occur for short sequences with 25 amino acids or less, and those with nonstandard amino acids. Errors are identified in an error report); (6) results: after a query is submitted, an icon 'Not Ready' appears in this field. Clicking on this icon opens the 'Job Info' box, which includes information about the query sequence, a link to the query results page, and an option to set a specific email address for that query; (7) saved-for: a countdown indicates the number of hours the query will be stored in the server. It can be reset to 48 h by clicking the refresh icon next to it.

Note that in addition to MoRFchibi$_{Light}$, the MoRFchibi server also computes MoRFchibi$_{Web}$, which relies on PSSM files generated by PSI-BLAST. The generation of PSSM files substantially increases the server processing time.

Accessing and reading results

● **TIMING 3 min per submitted protein sequence**

Once a job is completed, its runtime in seconds is shown in the status column, and the 'Not Ready' icon in the results field is replaced by two icons: 'Ready' and 'Graph' (Fig. 7b). The 'Ready' icon opens the 'Job Info' box, which at this point includes two extra icons, one to download the result in a text format and the other to open the graph window. The 'Graph' icon opens the graph window where the *x* axis is the amino acid index, and the *y* axis corresponds to the propensity scores of MoRFchibi$_{Light}$ (MCL) and its subcomponents, MoRFchibi (MC) and disorder prediction by Espritz-D (IDP) (Fig. 7b). Scores for MoRFchibi$_{Web}$ (MCW), MoRF$_{DC}$ (MDC) and the conservation scores (ICS) derived from PSSM files can also be visualized. The *y* axis is automatically bound to cover the range of predicted scores; however, users can change that range to (0, 1) by selecting the 'Toggle Y-axis Bound' checkbox. Moreover, the 'Toggle MoRF Bands' checkbox can be used to visualize the binary prediction of MoRFs. Users can select which scores to visualize by clicking on the corresponding name in the legend. They can also drag the mouse to select a region to zoom in, and the 'Reset zoom' can be used to zoom out. The graph can be downloaded in several formats from the print chart menu at the top right corner of the graph window. A text table with the results can be downloaded from three locations: the 'job Info' box, the results page, and the attachment of the 'results ready' email.

binary predictions, excluding arguably more informative real-valued propensities. Thus, we recommend flDPnn in scenarios when more accurate predictions are needed for small collections of proteins and IUPred when putative disorder propensities are needed for larger protein sets.

Prediction of disorder functions with ANCHOR2 and MoRFchibi

ANCHOR2 (ref. 112) offers reliable and fast prediction of IDRs that undergo disorder-to-order transition upon binding to a partner protein by relying on similar biophysical principles to IUPred[121]. As ANCHOR2 is available as an option on the IUPred web server site, the instructions are similar to the details in Box 2. The two main differences are to select 'ANCHOR2' from the 'Analysis type' box when submitting the query sequence (Fig. 6a), and the graphical panel for the results that includes two plots which represent the propensity values for the disordered binding generated by ANCHOR2 (in blue) and for disorder produced by IUPred (in red), Fig. 6b. The downloadable IUPred package contains the ANCHOR2 software, which can be executed as an optional flag without any further requirements.

MoRFchibi$_{Light}$ (ref. 114) is our recommended choice for the prediction of short disordered protein-binding regions. The predictions of MoRFchibi$_{Light}$ are assembled hierarchically using the Bayes rule in a stepwise fashion. First, a MoRFchibi prediction is generated by combining outputs of two support vector machine models that predict protein binding regions. Next, this result is combined with protein disorder prediction from Espritz[147] to separate binding segments in IDRs from those in structured regions. There are two ways to use this method, via the web server or as a downloadable software suite (both free for academic use):

- The web server version is at https://morf.msl.ubc.ca/index.xhtml. This version performs the entire prediction process on the MoRFchibi SYSTEM online server side, without the necessity to install any software. Box 3 and Fig. 7 describe the details of how to perform predictions and analyze the corresponding results using this web server
- Source code can be downloaded at https://gsponerlab.msl.ubc.ca/software/morf_chibi/downloads/. This version is available for Linux-based environments and must be installed and run on users' computers and requires the installation of Espritz[147] for disorder predictions. The input file 'input.fasta' can contain any number of sequences in a fasta format, and the result is saved in 'output.txt', which contains a table with the scores of MoRFchibi_light, MoRFchibi and rescaled ESpritz-D IDR predictions (for details, see ref. 114)

IDRs may interact with a variety of other ligands besides proteins and peptides that are covered by the predictions from MoRFchibi and ANCHOR2. These ligands include RNA, DNA, lipids, metals, ions, carbohydrates and small molecules[54,159–162]. There are relatively few methods

https://doi.org/10.1038/s41596-023-00876-x

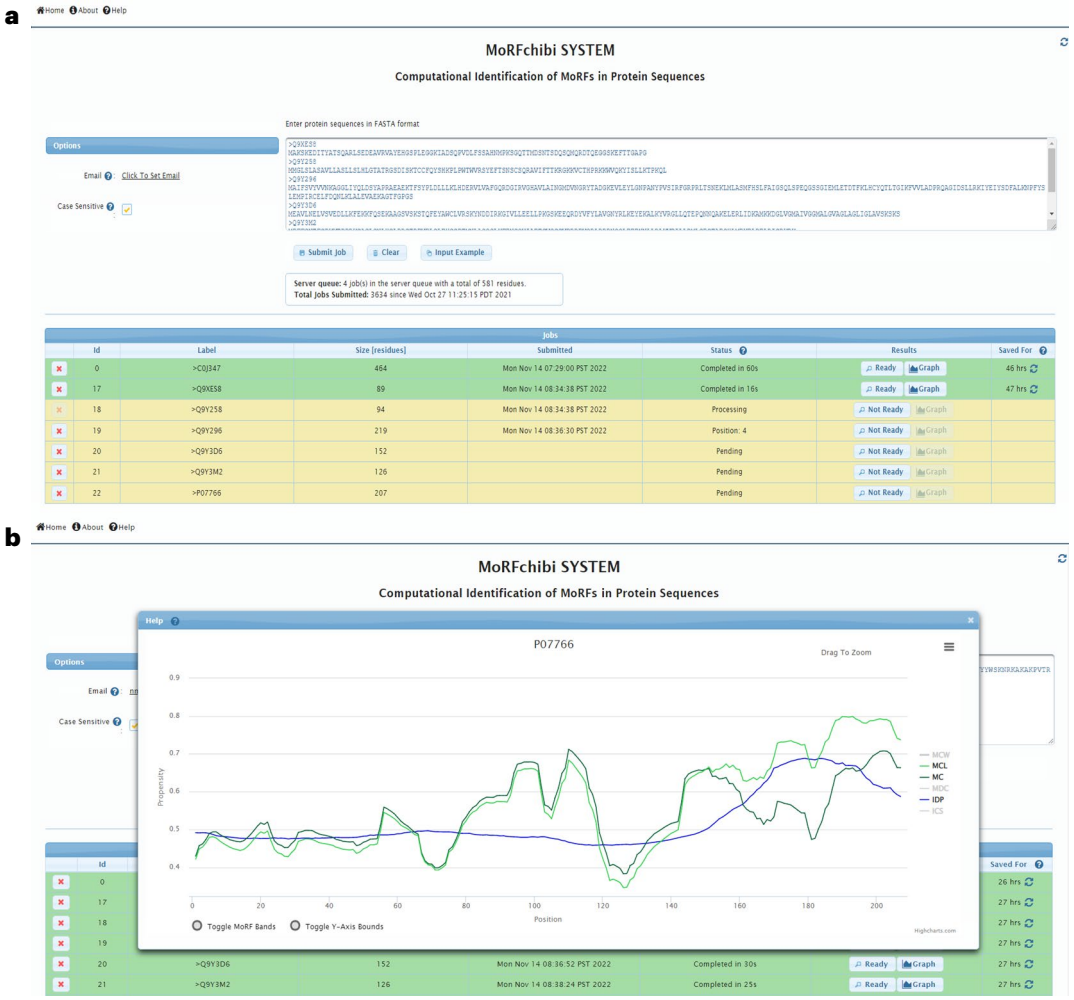

Fig. 7 | Example of MoRFchibi SYSTEM web server pages. a, The interface/input page where users submit their protein sequences. The page includes the jobs table where sequences that are waiting to be processed are highlighted in yellow and those that are completed are in green. The Q9Y258 sequence at the top of the server queue is being processed while Q9Y296 is fourth in the server queue. The last three sequences (Q9Y3D6, Q9Y3M2 and P07766) are in the user's private queue. **b**, The 'Graph' window for protein P07766 with scores of MoRFchibi$_{Light}$ (MCL in light green), basic MoRFchibi (MC in dark green) and disorder prediction by Espritz -D rescaled to fit a normal distribution (IDP in blue). Residues with higher values for MCL, MC or IDP imply a higher probability of being MoRFs, protein binding or disordered residues, respectively. Viewing of the MoRFchibi$_{Web}$ (MCW), MoRF$_{DC}$ (MDC) and conservation scores (ICS) is disabled.

that address predictions of these types of disordered binding regions and their performance was not yet tested in a community-driven experiment. While this precludes us from recommending specific tools, we highlight the availability of the DisorderEd PredictIon CenTER (DEPICTER) web server (http://biomine.cs.vcu.edu/servers/DEPICTER/) that integrates predictions of several types of disordered binding regions[163]. DEPICTER predicts disordered regions that interact with proteins using ANCHOR2 (ref. 112) and DisoRDPbind[113], regions that bind RNA and DNA using DisoRDPbind[113], and MoRFs using fMoRFpred[99]. It also predicts disordered linkers with DFLpred[132]. DisoRDPbind's model for the prediction of protein-binding regions secures an AUC of 0.729 and MCC of 0.198 with runtime <1 s, according to the CAID's results[70]. fMoRFpred is an older method that is only modestly accurate (AUC of 0.547 and MCC of 0.05 in CAID) but it generates results very quickly, with runtime <1 s per protein. Given the modest predictive performance of fMoRFpred, we suggest replacing its results with the predictions from MoRFchibi when using the DEPICTER resource.

Summary and future outlook

This guide for computational prediction of intrinsic disorder and disorder functions from protein sequences covers several predictive tools chosen for their accuracy, runtime, availability and their ability to predict disorder functions. We use results of the recently completed CAID experiment[70] to select methods that provide state-of-the-art and fast predictions. We also ensure that they are available to the end users in multiple modes, such as web servers and standalone code. We recommend two disorder predictors: flDPnn that is accurate and fast and IUPred that is very fast and moderately accurate. We also suggest two relatively accurate predictors of disordered binding IDRs: ANCHOR2, a very fast predictor of disordered binding regions, and MoRFchibi, a moderately fast method that predicts shorter MoRF regions. These methods provide complementary ways to conveniently obtain fast, state-of-the art predictions of intrinsic disorder and its functions. We describe how to find these methods, submit and collect their predictions, and read and interpret results that they generate. These details are provided by the authors of these tools, ensuring that the information is comprehensive. This covers several practical and often overlooked aspects, including limitations, options, timing and peculiarities of inputs and processing of predictions. We believe that this tutorial will help the end users in the selection of the right tools and will ease the learning curve on how to use and apply these methods.

Although current best-in-class disorder predictors, including those that we highlight, offer predictions that are accurate enough to be used in a practical context, their results should be considered with some

caution. A recent analysis that focuses on a predictive performance at the protein level, rather than typical evaluations that aggregate over datasets of proteins, reveals that while the majority of proteins are predicted accurately, up to 30% (depending on the predictor used) suffer relatively poor predictions[164]. These proteins typically have relatively high amounts of disorder. This finding is in line with other studies showing that disorder predictors produce less accurate results when predicting long IDRs[74] and have difficulty predicting fully disordered proteins[70]. Moreover, results from CAID reveal that predictors of disordered binding regions provide modest levels of predictive quality[70], suggesting that there is a lot of room for further improvements. One reason could be that the underlying binding annotations that are used to train and test these tools are fraught with more ambiguity than the disorder annotations. To be more specific, the exact position of binding regions is often inaccurate and binding annotations are typically extended into an entire IDR[70]. Another potential factor is that none of the binding predictors that participated in CAID utilize deep learning, while currently most accurate disorder predictors nearly exclusively depend on deep learning[65]. These disorder predictors rely on a variety of deep network architectures, including convolutional (AUCpred[81]), recurrent (SPOT-Disorder[165], IDP-Seq2Seq[166] and MetaPredict[167]), hybrids of convolutional and recurrent (SPOT-Disorder-Single[168], rawMSA[80], SPOT-Disorder2 (ref. 79) and RFPR-IDP[169]), and feed forward (flDPnn[78]). One option that is yet to be used to predict disorder or disorder function are transformer networks, which arguably improve over the above network types by applying attention mechanism and positional embeddings. The transformer networks were recently applied with success in related problems, including prediction of contact maps[170], protein–protein interactions[171] and protein–drug interactions[172]. Another relevant development are embeddings generated by the protein language models, which encode amino acids using numeric vectors that describe their surrounding sequence[173–175]. These embeddings were applied to a broad range of protein prediction problems, including a just-released disorder predictor, SETH[176]. Combining advanced network architectures, such as transformers, with modern sequence embeddings should lead to the development of more accurate disorder and disorder function predictors.

While IDRs interact with a broad range of partner molecules, current binding predictors are primarily focused on protein-binding IDRs[100,110]. Only a handful of methods target the prediction of nucleic acid binding IDRs (DeepDISObind[177] and DisoRDPbind[113,178]) and lipid-binding IDRs (DisoLipPred[179] and MemDis[180]). The development of tools for the prediction of interactions with other partner types, such as carbohydrates and metals, is currently infeasible due to an insufficient amount of ground truth data. However, as DisProt accumulates additional functional data, new predictors that address these functional types of IDRs are likely to be developed.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

References

1. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756 (2002).
2. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
3. Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **18**, 394–402 (2008).
4. van der Lee, R. et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
5. Oldfield, C. J., Uversky, V. N., Dunker, A. K. & Kurgan, L. in *Intrinsically Disordered Proteins* (ed. Salvi, N.) 1–34 (Academic Press, 2019).
6. Dunker, A. K. et al. What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* **1**, e24157 (2013).
7. Peng, Z. et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol. Life Sci.* **72**, 137–151 (2015).
8. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30**, 137–149 (2012).
9. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
10. Uversky, V. N. Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta* **1834**, 932–951 (2013).
11. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta* **1804**, 1231–1264 (2010).
12. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
13. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533 (2002).
14. Dunker, A. K. et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).
15. Tompa, P., Szasz, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **30**, 484–489 (2005).
16. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148 (2005).
17. Hu, G., Wu, Z., Uversky, V. N. & Kurgan, L. Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci.* **18**, 2761 (2017).
18. Patil, A., Kinoshita, K. & Nakamura, H. Domain distribution and intrinsic disorder in hubs in the human protein–protein interaction network. *Protein Sci.* **19**, 1461–1468 (2010).
19. Skinnider, M. A. et al. An atlas of protein–protein interactions across mouse tissues. *Cell* **184**, 4073–4089 e4017 (2021).
20. Holguin-Cruz, J. A., Foster, L. J. & Gsponer, J. Where protein structure and cell diversity meet. *Trends Cell Biol.* **32**, 996–1007 (2022).
21. Tantos, A., Han, K. H. & Tompa, P. Intrinsic disorder in cell signaling and gene transcription. *Mol. Cell Endocrinol.* **348**, 457–465 (2012).
22. Bondos, S. E., Dunker, A. K. & Uversky, V. N. Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Commun. Signal.* **20**, 20 (2022).
23. Darling, A. L. & Uversky, V. N. Intrinsic disorder and posttranslational modifications: the darker side of the biological dark matter. *Front. Genet.* **9**, 158 (2018).
24. Jakob, U., Kriwacki, R. & Uversky, V. N. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.* **114**, 6779–6805 (2014).
25. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384 (2005).
26. Trudeau, T. et al. Structure and intrinsic disorder in protein autoinhibition. *Structure* **21**, 332–341 (2013).
27. Buday, L. & Tompa, P. Functional classification of scaffold proteins and related molecules. *FEBS J.* **277**, 4348–4355 (2010).
28. Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in scaffold proteins: getting more from less. *Prog. Biophys. Mol. Biol.* **98**, 85–106 (2008).
29. Xue, B. et al. Stochastic machines as a colocalization mechanism for scaffold protein function. *FEBS Lett.* **587**, 1587–1591 (2013).
30. Fuxreiter, M. et al. Disordered proteinaceous machines. *Chem. Rev.* **114**, 6806–6843 (2014).

31. Romero, P. R. et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA* **103**, 8390–8395 (2006).
32. Zhou, J. H., Zhao, S. W. & Dunker, A. K. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *J. Mol. Biol.* **430**, 2342–2359 (2018).
33. Antifeeva, I. A. et al. Liquid–liquid phase separation as an organizing principle of intracellular space: overview of the evolution of the cell compartmentalization concept. *Cell. Mol. Life Sci.* **79**, 251 (2022).
34. Uversky, V. N. Recent developments in the field of intrinsically disordered proteins: Intrinsic disorder-based emergence in cellular biology in light of the physiological and pathological liquid-liquid phase transitions. *Annu. Rev. Biophys.* **50**, 135–156 (2021).
35. Uversky, V. N. Protein intrinsic disorder and structure–function continuum. *Prog. Mol. Biol. Transl. Sci.* **166**, 1–17 (2019).
36. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
37. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
38. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
39. Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2021).
40. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Sci.* **31**, e4353 (2022).
41. Baek, K. T. & Kepp, K. P. Assessment of AlphaFold2 for human proteins via residue solvent exposure. *J. Chem. Inf. Model.* **62**, 3391–3400 (2022).
42. Hemmings, H. C. Jr., Nairn, A. C., Aswad, D. W. & Greengard, P. DARPP-32, a dopamine- and adenosine 3′:5′-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. II. Purification and characterization of the phosphoprotein from bovine caudate nucleus. *J. Neurosci.* **4**, 99–110 (1984).
43. Gast, K. et al. Prothymosin alpha: a biologically active protein with random coil conformation. *Biochemistry* **34**, 13211–13218 (1995).
44. Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A. & Lansbury, P. T. Jr. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* **35**, 13709–13715 (1996).
45. Williams, R. M. et al. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* **2001**, 89–100 (2001).
46. Campen, A. et al. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15**, 956–963 (2008).
47. Zhao, B. & Kurgan, L. Compositional bias of intrinsically disordered proteins and regions and their predictions. *Biomolecules* **12**, 888 (2022).
48. Yan, J., Cheng, J., Kurgan, L. & Uversky, V. N. Structural and functional analysis of 'non-smelly' proteins. *Cell. Mol. Life Sci.* **77**, 2423–2440 (2020).
49. Romero, P. et al. Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).
50. Zhao, B. & Kurgan, L. Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Rev. Proteom.* **18**, 1019–1029 (2021).
51. Zhao, B. & Kurgan, L. in *Machine Learning in Bioinformatics of Protein Sequences* 205–236 (World Scientific, 2023).
52. Liu, Y., Wang, X. & Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **20**, 330–346 (2019).
53. Meng, F., Uversky, V. N. & Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.* **74**, 3069–3090 (2017).
54. Quaglia, F. et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* **50**, D480–D487 (2022).
55. Sickmeier, M. et al. DisProt: the database of disordered proteins. *Nucleic Acids Res.* **35**, D786–D793 (2007).
56. He, B. et al. Predicting intrinsic disorder in proteins: an overview. *Cell Res.* **19**, 929–949 (2009).
57. Meng, F., Uversky, V. & Kurgan, L. Computational prediction of intrinsic disorder in proteins. *Curr. Protoc. Protein Sci.* **88**, 2.16.11–12.16.14 (2017).
58. Deng, X., Eickholt, J. & Cheng, J. A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.* **8**, 114–121 (2012).
59. Dosztanyi, Z., Meszaros, B. & Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinform.* **11**, 225–243 (2010).
60. 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* **4**, e1259708 (2016).
61. Kurgan, L. Resources for computational prediction of intrinsic disorder in proteins. *Methods* **204**, 132–141 (2022).
62. Atkins, J. D., Boateng, S. Y., Sorensen, T. & McGuffin, L. J. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. *Int. J. Mol. Sci.* **16**, 19040–19054 (2015).
63. Williams, R. J. The conformation properties of proteins in solution. *Biol. Rev. Camb. Philos. Soc.* **54**, 389–437 (1979).
64. Eickholt, J. & Cheng, J. DNdisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics* **14**, 88 (2013).
65. Zhao, B. & Kurgan, L. Deep learning in prediction of intrinsic disorder in proteins. *Computat. Struct. Biotechnol. J.* **20**, 1286–1294 (2022).
66. Katuwawala, A. & Kurgan, L. Comparative assessment of intrinsic disorder predictions with a focus on protein and nucleic acid-binding proteins. *Biomolecules* **10**, 1636 (2020).
67. Necci, M., Piovesan, D., Dosztanyi, Z., Tompa, P. & Tosatto, S. C. E. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics* **34**, 445–452 (2018).
68. Peng, Z. L. & Kurgan, L. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.* **13**, 6–18 (2012).
69. Walsh, I. et al. Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* **31**, 201–208 (2015).
70. Necci, M., Piovesan, D., Predictors, C., DisProt, C. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481 (2021).
71. Jin, Y. & Dunbrack, R. L. Jr. Assessment of disorder predictions in CASP6. *Proteins* **61**, 167–175 (2005).
72. Bordoli, L., Kiefer, F. & Schwede, T. Assessment of disorder predictions in CASP7. *Proteins* **69**, 129–136 (2007).
73. Noivirt-Brik, O., Prilusky, J. & Sussman, J. L. Assessment of disorder predictions in CASP8. *Proteins* **77**, 210–216 (2009).
74. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A. & Fidelis, K. Assessment of protein disorder region predictions in CASP10. *Proteins* **82**, 127–137 (2014).
75. Melamud, E. & Moult, J. Evaluation of disorder predictions in CASP5. *Proteins* **53**, 561–565 (2003).
76. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A. & Kryshtafovych, A. Evaluation of disorder predictions in CASP9. *Proteins* **79**, 107–118 (2011).
77. Lang, B. & Babu, M. M. A community effort to bring structure to disorder. *Nat. Methods* **18**, 454–455 (2021).

78. Hu, G. et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 4438 (2021).
79. Hanson, J., Paliwal, K. K., Litfin, T. & Zhou, Y. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteom. Bioinforma.* **17**, 645–656 (2019).
80. Mirabello, C. & Wallner, B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PLoS ONE* **14**, e0220182 (2019).
81. Wang, S., Ma, J. & Xu, J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* **32**, i672–i679 (2016).
82. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
83. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
84. Kim, S. S., Seffernick, J. T. & Lindert, S. Accurately predicting disordered regions of proteins using rosetta residuedisorder application. *J. Phys. Chem. B* **122**, 3920–3930 (2018).
85. He, J., Turzo, S. B. A., Seffernick, J. T., Kim, S. S. & Lindert, S. Prediction of intrinsic disorder using rosetta residuedisorder and AlphaFold2. *J. Phys. Chem. B* **126**, 8439–8446 (2022).
86. Wilson, C. J., Choy, W. Y. & Karttunen, M. AlphaFold2: a role for disordered protein/region prediction? *Int. J .Mol. Sci.* **23**, 4591 (2022).
87. Piovesan, D., Monzon, A. M. & Tosatto, S. C. E. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci.* **31**, e4466 (2022).
88. Aderinwale, T. et al. Real-time structure search and structure classification for AlphaFold protein models. *Commun. Biol.* **5**, 316 (2022).
89. Kurgan, L., Li, M. & Li, Y. in *Systems Medicine* (ed. Wolkenhauer, O.) 159–169 (Academic Press, 2021).
90. Zhao, B. et al. Intrinsic disorder in human RNA-binding proteins. *J. Mol. Biol.* **433**, 167229 (2021).
91. Zhao, B., Katuwawala, A., Uversky, V. N. & Kurgan, L. IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell. *Cell. Mol. Life Sci.* **78**, 2371–2385 (2020).
92. Giri, R. et al. Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses. *Cell. Mol. Life Sci.* **78**, 1655–1688 (2020).
93. Cubuk, J. et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* **12**, 1936 (2021).
94. Kumar, N. et al. Comprehensive intrinsic disorder analysis of 6108 viral proteomes: from the extent of intrinsic disorder penetrance to functional annotation of disordered viral proteins. *J. Proteome Res.* **20**, 2704–2713 (2021).
95. Zou, H. et al. Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes. *Nucleic Acids Res.* **50**, e49 (2022).
96. Meszaros, B., Hajdu-Soltesz, B., Zeke, A. & Dosztanyi, Z. Mutations of intrinsically disordered protein regions can drive cancer but lack therapeutic strategies. *Biomolecules* **11**, 381 (2021).
97. Oldfield, C. J. et al. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12454–12470 (2005).
98. Vacic, V. et al. Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* **6**, 2351–2366 (2007).
99. Yan, J., Dunker, A. K., Uversky, V. N. & Kurgan, L. Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* **12**, 697–710 (2016).
100. Katuwawala, A., Peng, Z. L., Yang, J. Y. & Kurgan, L. Computational prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput. Struct. Biotechnol. J.* **17**, 454–462 (2019).
101. Mohan, A. et al. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059 (2006).
102. Oldfield, C. J. et al. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9**, S1 (2008).
103. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384 (2005).
104. Uversky, V. N. Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chem. Soc. Rev.* **40**, 1623–1634 (2011).
105. Fuxreiter, M. Fuzzy protein theory for disordered proteins. *Biochem. Soc. Trans.* **48**, 2557–2564 (2020).
106. Miskei, M. et al. Fuzziness enables context dependence of protein interactions. *FEBS Lett.* **591**, 2682–2695 (2017).
107. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
108. Berlow, R. B., Dyson, H. J. & Wright, P. E. Multivalency enables unidirectional switch-like competition between intrinsically disordered proteins. *Proc. Natl Acad. Sci. USA* **119**, e2117338119 (2022).
109. Bhowmick, P., Guharoy, M. & Tompa, P. Bioinformatics approaches for predicting disordered protein motifs. *Adv. Exp. Med. Biol.* **870**, 291–318 (2015).
110. Katuwawala, A., Ghadermarzi, S. & Kurgan, L. Computational prediction of functions of intrinsically disordered regions. *Prog. Mol. Biol. Transl. Sci.* **166**, 341–369 (2019).
111. Basu, S., Kihara, D. & Kurgan, L. Computational prediction of disordered binding regions. *Comput. Struct. Biotechnol. J.* **21**, 1487–1497 (2023).
112. Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
113. Peng, Z. & Kurgan, L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* **43**, e121 (2015).
114. Malhis, N., Jacobson, M. & Gsponer, J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* **44**, W488–W493 (2016).
115. Malhis, N., Wong, E. T., Nassar, R. & Gsponer, J. Computational identification of MoRFs in protein sequences using hierarchical application of Bayes rule. *PLoS ONE* **10**, e0141603 (2015).
116. Sharma, R., Raicar, G., Tsunoda, T., Patil, A. & Sharma, A. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics* **34**, 1850–1858 (2018).
117. Uversky, V. N. p53 proteoforms and intrinsic disorder: an illustration of the protein structure-function continuum concept. *Int. J. Mol. Sci.* **17**, 1874 (2016).
118. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
119. Lewis, T. E. et al. Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **46**, D435–D439 (2018).
120. Schad, E. et al. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**, 535–537 (2018).
121. Erdos, G., Pajkos, M. & Dosztanyi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).
122. Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**, 1402–1404 (2017).

123. Kozlowski, L. P. & Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* **13**, 111 (2012).
124. Fan, X. & Kurgan, L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. *J. Biomol. Struct. Dyn.* **32**, 448–464 (2014).
125. Anderson, C. W. & Appella, E. in *Handbook of Cell Signaling* (eds, Bradshaw, R. A. & Dennis, E. A.) 237–247 (Academic Press, 2004).
126. Campbell, S. J., Edwards, R. A. & Glover, J. N. Comparison of the structures and peptide binding specificities of the BRCT domains of MDC1 and BRCA1. *Structure* **18**, 167–176 (2010).
127. Christou, C. M. & Kyriacou, K. BRCA1 and its network of interacting partners. *Biology* **2**, 40–63 (2013).
128. Mark, W. Y. et al. Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein–protein and protein–DNA interactions? *J. Mol. Biol.* **345**, 275–287 (2005).
129. Deng, C. X. & Brodie, S. G. Roles of BRCA1 and its interacting proteins. *Bioessays* **22**, 728–737 (2000).
130. Dosztanyi, Z. Prediction of protein disorder based on IUPred. *Protein Sci.* **27**, 331–340 (2018).
131. Peng, Z., Wang, C., Uversky, V. N. & Kurgan, L. Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol. Biol.* **1484**, 187–203 (2017).
132. Meng, F. & Kurgan, L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* **32**, i341–i350 (2016).
133. Buchan, D. W. A. & Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
134. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
135. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
136. Oates, M. E. et al. D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* **41**, D508–D516 (2013).
137. Piovesan, D. et al. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367 (2021).
138. Potenza, E., Di Domenico, T., Walsh, I. & Tosatto, S. C. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* **43**, D315–D320 (2015).
139. Di Domenico, T., Walsh, I., Martin, A. J. M. & Tosatto, S. C. E. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* **28**, 2080–2081 (2012).
140. Piovesan, D. et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* **46**, D471–D476 (2018).
141. Zhao, B. et al. DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* **49**, D298–D308 (2021).
142. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208 (2006).
143. Romero, P. et al. Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).
144. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).
145. Ishida, T. & Kinoshita, K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* **24**, 1344–1348 (2008).
146. Ghalwash, M. F., Dunker, A. K. & Obradovic, Z. Uncertainty analysis in protein disorder prediction. *Mol. Biosyst.* **8**, 381–391 (2012).
147. Walsh, I., Martin, A. J., Di Domenico, T. & Tosatto, S. C. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
148. Linding, R. et al. Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
149. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).
150. Monzon, A. M., Rohr, C. O., Fornasari, M. S. & Parisi, G. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database* https://doi.org/10.1093/database/baw038 (2016).
151. Hatos, A. et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276 (2020).
152. Dinkel, H. et al. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **44**, D294–D300 (2016).
153. Miskei, M., Antal, C. & Fuxreiter, M. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.* **45**, D228–D235 (2017).
154. Fukuchi, S. et al. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* **42**, D320–D325 (2014).
155. Ficho, E., Remenyi, I., Simon, I. & Meszaros, B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **33**, 3682–3684 (2017).
156. consortium, P. D.-K. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* **50**, D534–D542 (2022).
157. Meszaros, B. et al. PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res* **48**, D360–D367 (2020).
158. UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
159. Kjaergaard, M. & Kragelund, B. B. Functions of intrinsic disorder in transmembrane proteins. *Cell. Mol. Life Sci.* **74**, 3205–3224 (2017).
160. Wu, Z. H. 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.* **589**, 2561–2569 (2015).
161. Chowdhury, S., Zhang, J. & Kurgan, L. In silico prediction and validation of novel RNA binding proteins and residues in the human proteome. *Proteomics* **18**, e1800064 (2018).
162. Wang, C., Uversky, V. N. & Kurgan, L. Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* **16**, 1486–1498 (2016).
163. Barik, A. et al. DEPICTER: intrinsic disorder and disorder function prediction server. *J. Mol. Biol.* **432**, 3379–3387 (2020).
164. Katuwawala, A., Oldfield, C. J. & Kurgan, L. Accuracy of protein-level disorder predictions. *Brief. Bioinform.* **21**, 1509–1522 (2020).
165. Hanson, J., Yang, Y., Paliwal, K. & Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685–692 (2017).
166. Tang, Y. J., Pang, Y. H. & Liu, B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics* **36**, 5177–5186 (2021).
167. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120**, 4312–4319 (2021).
168. Hanson, J., Paliwal, K. & Zhou, Y. Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures. *J. Chem. Inf. Model.* **58**, 2369–2376 (2018).

169. Liu, Y., Wang, X. & Liu, B. 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.* **22**, 2000–2011 (2021).
170. Singh, J., Litfin, T., Singh, J., Paliwal, K. & Zhou, Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **38**, 1888–1894 (2022).
171. Ieremie, I., Ewing, R. M. & Niranjan, M. TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. *Bioinformatics* **38**, 2269–2277 (2022).
172. Yan, X. & Liu, Y. Graph-sequence attention and transformer for predicting drug-target affinity. *RSC Adv.* **12**, 29525–29534 (2022).
173. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
174. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
175. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669 e653 (2021).
176. Ilzhofer, D., Heinzinger, M. & Rost, B. SETH predicts nuances of residue disorder from protein embeddings. *Front. Bioinform.* **2**, 1019597 (2022).
177. Zhang, F., Zhao, B., Shi, W., Li, M. & Kurgan, L. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. *Brief. Bioinform.* **23**, bbab521 (2022).
178. Peng, Z. L., Wang, C., Uversky, V. N. & Kurgan, L. Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol. Biol.* **1484**, 187–203 (2017).
179. Katuwawala, A., Zhao, B. & Kurgan, L. DisoLipPred: accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning. *Bioinformatics* **38**, 115–124 (2021).
180. Dobson, L. & Tusnady, G. E. MemDis: predicting disordered regions in transmembrane proteins. *Int. J. Mol. Sci.* **22**, 12270 (2021).
181. Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* **22**, 2948–2949 (2006).
182. Lobanov, M. Y. & Galzitskaya, O. V. The Ising model for prediction of disordered residues from protein sequence alone. *Phys. Biol.* **8**, 035004 (2011).
183. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
184. Iqbal, S. & Hoque, M. T. DisPredict: a predictor of disordered protein using optimized RBF Kernel. *PLoS ONE* **10**, e0141551 (2015).
185. Orlando, G., Raimondi, D., Codice, F., Tabaro, F. & Vranken, W. Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J. Mol. Biol.* **434**, 167579 (2022).
186. Kumar, M. et al. ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **48**, D296–D306 (2020).
187. 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.* **40**, D261–D270 (2012).
188. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
189. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Acknowledgements

This research was supported in part by the National Science Foundation (2146027 and 2125218 to L.K.), Robert J. Mattauch Endowment funds to L.K., and National Natural Science Foundation of China (31970649 to G.H. and K.W.). Z.D. acknowledges funding from the European Union's Horizon 2020 research and innovation programme (778247) and support from ELIXIR Hungary (www.elixir-hungary.org) and ELIXIR Implementations Studies. E.G. acknowledges support from the ÚNKP-22-1 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. J.G. acknowledges support from the National Sciences and Engineering Research Council of Canada.

Author contributions

L.K. conceptualized and coordinated the study, and collected and analyzed data. G.H., K.W. S.G., B.Z. and L.K. contributed to the development and to the description of the flDPnn tool. N.M. and J.G. contributed to the development and the description of the MoRFchibi tool. G.E. and Z.D. contributed to the development and the description of the IUPred and ANCHOR tools. V.N.U. contributed to formulation and discussion of the examples. L.K., V.N.U. and Z.D. contributed to writing the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-023-00876-x.

Correspondence and requests for materials should be addressed to Lukasz Kurgan, Jörg Gsponer, Vladimir N. Uversky or Zsuzsanna Dosztányi.

Peer review information *Nature Protocols* thanks the anonymous reviewers for their contribution to the peer review of this work.

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA. [2]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China. [3]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [4]MTA-ELTE Momentum Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary. [5]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. [6]Byrd Alzheimer's Center and Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA.

nature portfolio

Corresponding author(s): Lukasz Kurgan

Last updated by author(s): 2023-06-06

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	not applicable
Data analysis	not applicable

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data supporting the findings of this study are available within the paper and its Supplementary Information.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

(●) Life sciences ( ) Behavioural & social sciences ( ) Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	not applicable
Data exclusions	not applicable
Replication	not applicable
Randomization	not applicable
Blinding	not applicable

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing	
Data exclusions	
Non-participation	
Randomization	

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	

Blinding

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

Location

Access & import/export

Disturbance

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
	Antibodies
	Eukaryotic cell lines
	Palaeontology and archaeology
	Animals and other organisms
	Clinical data
	Dual use research of concern

Methods

n/a	Involved in the study
	ChIP-seq
	Flow cytometry
	MRI-based neuroimaging

Antibodies

Antibodies used

Validation

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines
(See ICLAC register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Dating methods

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Study protocol

Data collection

Outcomes

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
○	○	Public health
○	○	National security
○	○	Crops and/or livestock
○	○	Ecosystems
○	○	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
○	○	Demonstrate how to render a vaccine ineffective
○	○	Confer resistance to therapeutically useful antibiotics or antiviral agents
○	○	Enhance the virulence of a pathogen or render a nonpathogen virulent
○	○	Increase transmissibility of a pathogen
○	○	Alter the host range of a pathogen
○	○	Enable evasion of diagnostic/detection modalities
○	○	Enable the weaponization of a biological agent or toxin
○	○	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

Files in database submission

Genome browser session

(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Acquisition

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI ○ Used ○ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ○ Whole brain ○ ROI-based ○ Both

Statistic type for inference (See Eklund et al. 2016)

Correction

Models & analysis

n/a | Involved in the study

- Functional and/or effective connectivity
- Graph analysis
- Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis