www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Quality and bias of protein disorder predictors

Jakob T. Nielsen[1,2] & Frans A. A. Mulder[1,2]

Received: 4 December 2018
Accepted: 13 March 2019
Published online: 26 March 2019

Disorder in proteins is vital for biological function, yet it is challenging to characterize. Therefore, methods for predicting protein disorder from sequence are fundamental. Currently, predictors are trained and evaluated using data from X-ray structures or from various biochemical or spectroscopic data. However, the prediction accuracy of disordered predictors is not calibrated, nor is it established whether predictors are intrinsically biased towards one of the extremes of the order-disorder axis. We therefore generated and validated a comprehensive experimental benchmarking set of site-specific and continuous disorder, using deposited NMR chemical shift data. This novel experimental data collection is fully appropriate and represents the full spectrum of disorder. We subsequently analyzed the performance of 26 widely-used disorder prediction methods and found that these vary noticeably. At the same time, a distinct bias for over-predicting order was identified for some algorithms. Our analysis has important implications for the validity and the interpretation of protein disorder, as utilized, for example, in assessing the content of disorder in proteomes.

Interest in intrinsically disordered proteins (IDPs) has grown immensely over the past decades. IDPs can serve a large range of functions due to their enhanced sampling of conformational space compared to structured proteins and their involvement in many important biological processes and diseases have been discovered recently[1–7]. Although experimental characterization of IDPs is very challenging, protein sequence composition has distinct biases and this has inspired the development of a large number of computational methods for predicting disorder from sequence[8,9]. Recently, predictions of disorder by various methods have been compiled into databases[10–12] enabling consensus predictions, and meta-methods have emerged that predict disorder based on output from other predictors[13–15]. Protein disordered region (DR) prediction has been assessed periodically through the critical assessment of structure prediction (CASP) initiative[16]. DR predictions did not improve from CASP8 to CASP9[17], and only slightly for CASP10[18]. This apparent stagnation in accuracy of disorder predictors would suggest that development of new more sophisticated predictors would not have sufficient merit, and DR predictions were not evaluated anymore in subsequent CASP assessments.

We argue that this stagnation can be attributed to the vague authority of the evaluation, caused by insufficient quality of the data used to evaluate (and train) the predictors: In CASP, DR predictors were evaluated using missing density in X-ray structures as the disorder criterion. However, regions in X-ray structures might falsely appear ordered due to biases in non-native conditions required for X-ray crystallography characterization. In addition, since only proteins amenable to X-ray diffraction are included, such data sets are imbalanced in the sense that missing residues are relatively rare (only 2.4% in the set analyzed here) causing balance problems in the model building. As a complement, disorder analysis can be done for proteins in solution, as done in the DisProt database[19,20], and this data collection has frequently been used to train and evaluate disorder predictors[21]. Unfortunately, DisProt suffers from a heterogeneous compilation of data from diverse experimental sources, such as CD and sensitivity to proteolytic degradation, which lack position-specific information. Several false positive IDPs were indeed found in DisProt in a previous analysis[22]. A more serious issue arises from the fact that all currently applied evaluation criteria are binary classifiers, which ignore meaningful, intermediate order or a continuous range of structure[1,23,24], and therewith limit disorder prediction to a low-precision binary-classification problem. A more balanced dataset with a higher precision and accuracy would renew the potential in the development of bioinformatics methods for predicting disorder from sequence. For this purpose, we resorted to experimental data from NMR spectroscopy.

[1]Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Gustav Wieds Vej 14, 8000, Aarhus C, Denmark. [2]Department of Chemistry, Aarhus University, Langelandsgade 140, 8000, Aarhus C, Denmark. Correspondence and requests for materials should be addressed to J.T.N. (email: jtn@inano.au.dk) or F.A.A.M. (email: fmulder@chem.au.dk)

It is well-established that proteins can be studied with high accuracy in solution under near-native conditions by NMR spectroscopy. First, the structure-determination process provides an ensemble of structures where each model is consistent with the experimental data[25–29]. Second, and more quantitatively, nuclear spin relaxation rates provide information about the time-scale and amplitude of dynamics in proteins[30–32], capturing and validating the variability in the NMR structures. Unfortunately, spin relaxation experiments and data analysis are relatively complicated to pursue, are not applicable for all time scales or for IDPs, and therefore there is very little data available for highly dynamic sites in proteins[33]. Thirdly, chemical shifts are very sensitive to the local structure, are measured routinely and with very high precision for both structured proteins and IDPs[34,35], and have been used extensively to report on protein structure and dynamics[36,37]. In particular, chemical shifts and their deviation from random coil values have been used to determine and quantify order/disorder and conformational propensities in IDPs[38–43]. Modern molecular dynamics (MD) simulations reproduce experimental dynamical data with increasing accuracy[44] and, in particular, spin relaxation data has been used as an exquisite standard to benchmark MD force fields[45]. IDPs can be simulated with high accuracy in the description of local conformational equilibria, and a very close agreement has been established between the degree of order/disorder in IDPs and secondary chemical shifts[46–48].

Recently, we introduced the Chemical shift Z-score for assessing Order/Disorder (the CheZOD score)[22], which is based on deviations from random coil chemical shifts (RCCSs) using our refined formulation of RCCS reference values[49]. In contrast to other methods for describing order/disorder, this CheZOD Z-score provides a position-specific and continuous measure of order/disorder in proteins. Furthermore, the corresponding CheZOD database of such Z-scores for 117 proteins studied at near-native conditions is diverse and balanced, containing equal amounts of disordered and ordered residues[22]. Here, we rigorously benchmark the performance of 26 disorder prediction methods by assessing the agreement between the estimated probabilities of disorder and the experimental Z-scores for each predictor, and use this result to rank the accuracy of the predictors. We observed that the accuracy of the predictors depends on the type of features applied, the method of optimization, and that the newest predictors are generally the most accurate. Some predictors are biased towards over-predicting order. Our analysis suggests that current DR predictions are limited by the quality of the training data rather than by the capacity of the data mining approaches. Improved predictors can therefore be anticipated.

Results

Measures of disorder and flexibility in protein structures: p53 as an example.

To illustrate the process of disorder assignment, we consider the human oncogene protein p53, which contains ordered as well as disordered domains and is often used for illustrating predictions of disorder and interactions in IDPs[50,51]. p53 is interesting because of its involvement in more than 50% of human cancers and many diverse biological processes due to its multitude of conformations[46–48]. Estimated disorder probabilities for a large number of prediction methods (Fig. 1a, obtained from the genesilico server[13]) show agreement for some regions, but also substantial differences between the individual predictors. It is not possible to identify the most appropriate predictor *a priori* although that choice would have a dramatic impact for the prediction of disordered regions (see Supplementary Fig. S1 for prediction examples for 5 additional proteins). Consensus predictions from MobiDB-lite[11] (Fig. 1b) and D^2P^2[10] (Fig. 1c) suggest disorder outside of the structured domains and higher probability of disorder for the loops in the core domain (e.g. res. 181–191). However, disorder is also predicted for part of a rigid internal beta-strand in the core domain (res. 156–162) and for the entire folded tetramerization domain. When the DisProt database[20] (Fig. 1d) is used to assign disorder, two loop regions are assigned as *confident* disorder (res. 114–120 and 182–187), whereas the linker between the core domain and tetramerization domain (res. 293–312) shows *ambiguous* disorder. The remaining residues are classified as *context-dependent*, meaning that these regions cannot be assigned unequivocally to a disordered/ordered state. X-ray structures for the p53 core domain have missing densities for the ends of some of the sequence constructs. In contrast, internal residues with missing densities were only observed for two of the 12 chains for the loop comprising residues Lys120 and Ser121, which were also classified confidently as disordered in the DisProt database (Fig. 1e). A continuous measure for local disorder/order, for which data is more abundant and balanced, is the local structural variation in an NMR ensemble. Here we introduce two types of structural order parameters, *S* and *T*, based on NMR ensemble variation in dihedral angles and the Cα internal distances, respectively, (see Online Methods). These order parameters span from zero to unity, ranging from complete disorder to order, and are in qualitative agreement with disorder predictions (e.g. for the two confident DisProt disorder regions) and show dips in order/increase in flexibility in all the loop regions of the core domain (Fig. 1h). Finally, we provide experimental disorder through the introduction of a continuous site-specific descriptor derived from assigned chemical shifts[22] for p53[52] (see Fig. 1i). According to these Z-scores, the core domain and tetramerization domain are ordered, whereas several loops in the core domain are disordered to varying degree (Fig. 1i). For example, the loop comprising residues Lys120 and Ser121. There is a very close agreement between disorder from Z-scores and structural flexibility in the NMR ensemble (Fig. 1f,g). A more comprehensive systematic comparison, for a large set of proteins, reveals good agreement between CheZOD Z-scores and other measures of disorder, including structural variability in MD simulations[53,54] (see Supplementary Results 1 and Supplementary Figs S2–S5).

Benchmarking the performance of disorder predictors.

Above, a qualitative agreement was observed between Z-scores and estimated disorder probabilities for p53 with some noteworthy differences between individual predictors. To analyze the agreement systematically, disorder predictions were obtained for the 117 proteins in the CheZOD database as described in Online Methods. The calculated Z-scores were compared to the estimated disorder probabilities for a large set of different disorder predictors (see Table 1 and Online Methods) with the aim of identifying the best methods as those having the best agreement between estimated disorder probabilities and Z-scores. Figure 2 shows scatter plots of the Z-scores vs. the estimated probabilities (Z vs. p)

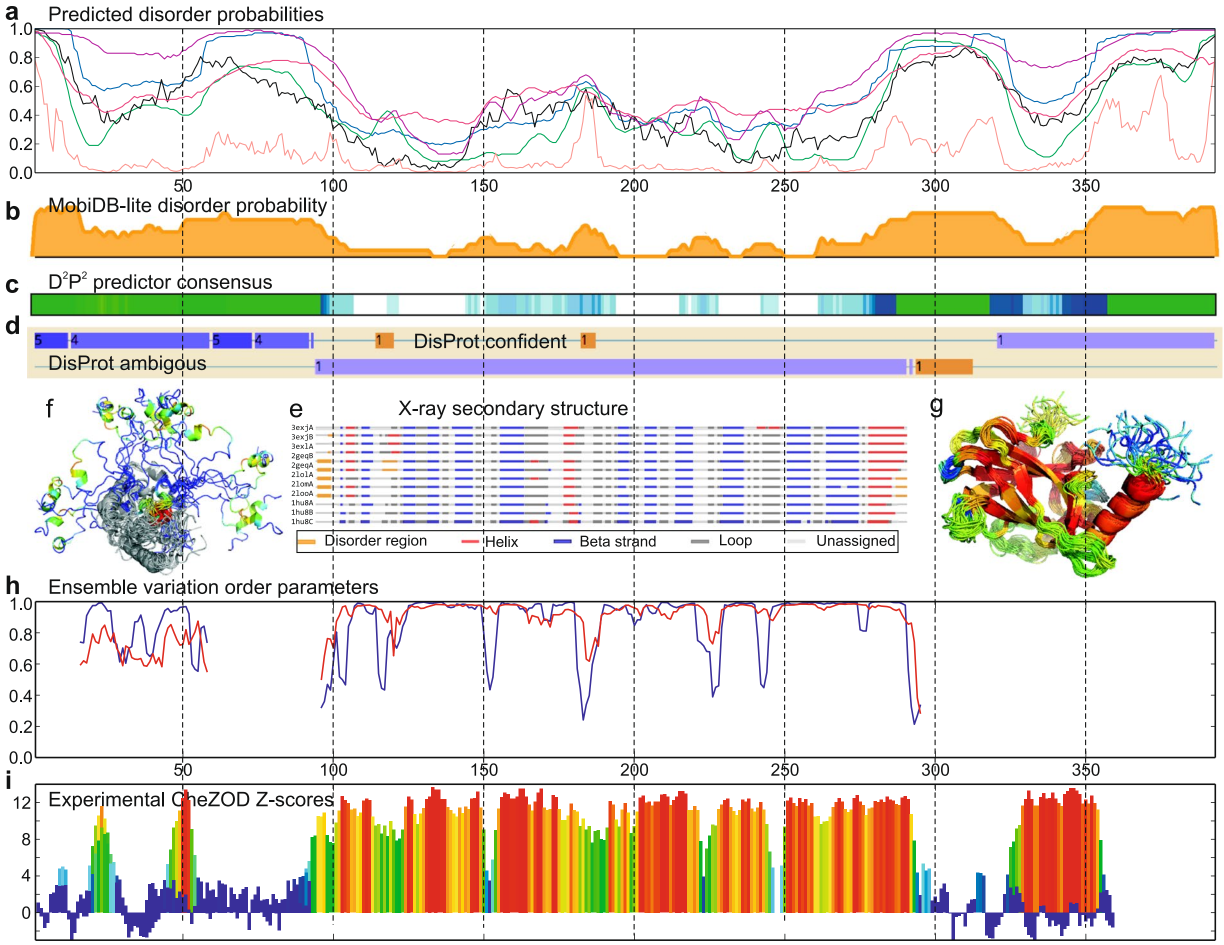

Figure 1. p53 experimental and inferred disorder. (**a**) Predicted disorder for PrDOS (green), IUPred_short (black), MetaDisorderMD2 (blue), RONN (pink), DISPROT_(VSL2b) (purple) and DISOPRED2 (light red). (**b**) Disorder probabilities from MobiDB-lite. (**c**) Agreement between disorder predictions from the D^2P^2 database shown as color intensity in a gradient bar. The green bars encode predicted disorder in segments outside predicted SCOP domains. The blue segments are where the disorder predictions intersect the SCOP domain prediction. (**d**) Inferred disorder/order from DisProt showing" disorder" and "context-dependent" regions with light brown and purple, respectively. (**e**) Assigned secondary structures for aligned chains with at least 95% sequence identity to 2FEJ analyzed and displayed using the PDBFlex server[75]. (**f**,**g**) NMR ensemble structure for the core domain (**g**) and N-terminal (**f**) as above colored according to CheZOD Z-scores as in (**i**). (**h**) NMR ensemble variation for p53 core domain (pdb id 2FEJ)[76] and N-terminal residues 14–60 bound to HMGB1[77] (pdb id 2lya, chain B) using red and blue lines for coordinate and angle order parameters, respectively, (see Online Methods). See also Supplementary Fig. S5 for more Z-score/flexibility protein profiles. (**i**) Experimental CheZOD Z-scores for p53 using previously assigned chemical shifts for res. 82–360[52], data for N-term res. 1–92 from Fersht *et al.*[78] and res. 14–60 bound to HMGB1 (in the background) as in (**g**) are shown superimposed. Z-scores are displayed with bars colored from blue through green through yellow to red indicating the highest scores corresponding to ordered residues.

for each predictor. It is seen that most predictors provide relatively high estimated probabilities of disorder for residues with low Z-scores and correspondingly lower probabilities for residues with high Z-scores. Qualitative agreement is observed, but the predictions are clearly different, with different qualities and biases in the correlation with Z-scores. To assess this agreement quantitively, we take full advantage of the continuous descriptor of disorder by determining the Pearson correlation coefficient, R_P, of agreement (see Fig. 3). This number is ideal for ranking the predictors from the best (largest absolute value) to the worst. As Z-scores increase with *order* while *p* is a measure of *disorder*, −1 indicates a perfect correlation and 0 expresses a complete lack of correlation. It is seen that binary predictors show poor correlation, while the newer, continuous methods SPOT-disorder[55], MFDp2[14] and AUCpreD[56] predict best (Table 1 and Fig. 3). Furthermore, the genesilico metapredictors[13] perform slightly better than all the methods used by the metapredictors but slightly inferior to the newer methods mentioned above (Table 1 and Fig. 3). The ESpritz[57] methods perform increasingly well, when trained on DisProt data, X-ray data, and NMR data, respectively (Table 1 and Fig. 3). Two methods that use NMR data for training – s2D[58] and DynaMine[59] – were also included. These methods were trained on continuous-valued target data; i.e. chemical shift derived secondary structure populations for *s2D* and local fast dynamics, as defined by the order parameter,

Method	R_P	R_S	AUC	pZA	pZD	Input[a]	Class[b]
MFDp2	−0.631	−0.592	0.853	0.582	0.490	Pdis	Meta
MetaDisorderMD2	−0.614	−0.579	0.852	0.513	0.325	Pdis	Meta
MetaDisorderMD	−0.616	−0.580	0.853	0.479	0.308	Pdis	Meta
MetaDisorder	−0.617	−0.575	0.865	0.590	0.399	Pdis	Meta
MetaDisorder3D	−0.361	−0.352	0.727	0.261	0.126	ST	ML
SPOT-dis	−0.657	−0.638	0.881	0.426	0.475	Evo	ML
AUCpreD	−0.598	−0.588	0.865	0.441	0.552	Evo	ML
PrDOS	−0.541	−0.543	0.836	0.403	0.277	Evo/ST	ML
RONN	−0.500	−0.495	0.804	0.525	0.172	Evo	ML
DISpro	−0.437	−0.498	0.805	0.221	0.269	Evo	ML
DISOPRED2	−0.330	−0.404	0.738	0.120	0.109	Evo	ML
DISOPRED3	−0.551	−0.553	0.833	0.332	0.388	Evo	ML
s2D[c]	−0.528	−0.501	0.797	0.610	0.241	Evo	ML
Dynamine[c]	−0.505	−0.489	0.806	0.502	0.124	AA	ML
ESpritz_NMR	−0.478	−0.483	0.797	0.335	0.300	AA	ML
ESpritz_Xray	−0.438	−0.474	0.791	0.208	0.230	AA	ML
ESpritz_DisProt	−0.419	−0.374	0.748	0.575	0.209	AA	ML
AUCpreD_noEvo	−0.512	−0.552	0.841	0.386	0.460	AA	ML
DISPROT (VSL2b)	−0.536	−0.497	0.808	0.609	0.286	AA	ML
IUPred_long	−0.566	−0.541	0.834	0.493	0.302	AA	SF
IUPred_short	−0.532	−0.505	0.822	0.424	0.275	AA	SF
Pdisorder[d]	−0.480	n.a.	n.a.	0.523	0.430	AA	ML
DisEMBL_coils	−0.404	−0.364	0.735	0.523	0.150	AA	ML
DisEMBL_remark465	−0.386	−0.389	0.737	0.405	0.140	AA	ML
DisEMBL_hotloops	−0.286	−0.334	0.702	0.109	0.049	AA	ML
GlobPlot[d]	−0.014	n.a.	n.a.	0.064	0.008	AA	SF

Table 1. Performance of disorder predictors. [a]Input: AA (AA type/property and composition); Evo (evolutionary information based on multiple sequence alignment profiles); ST (Structural templates); Pdis (estimated disorder probabilities from other predictors). [b]Class: Meta (meta-predictor); ML (machine learning); SF (scoring function). [c]Predicts continuous-valued NMR parameters (see Methods). Since the prediction output is not an actual disorder probability, the derivation of pZA and pZH do not strictly apply (see text and Online Methods). [d]Binary prediction methods. Derivation of Spearman correlation and AUC do not apply.

for DynaMine. Here we interpret the predicted populations of non-alpha-helix/beta-sheet as the probability of disorder and use a bijective transformation of the predicted order parameters to convert it to a pseudo-probability (see Online Methods). Judged by the Pearson correlation coefficient, these two methods are ranked in the middle for predicting Z-scores. The Spearman rank correlation coefficient, R_S, describing the agreement with a monotonic relationship between p and Z (not necessarily linear) was also calculated, and showed the same trend for the predictors (see Table 1 and Supplementary Fig. S6).

It is evident from Fig. 2 that predictions and Z-scores cluster in four quadrants due to the underlying bimodal distribution of Z-scores[22] and the binary nature of the classification used for training the methods. A very slight over-representation of "medium-range" Z-scores (close to 8.0) for average probabilities (close to 0.5) is seen only for the best ranked methods and IUPred[60]. To enable comparison with previous benchmarks, we also performed analysis for a binary classification of disorder using the definition $Z < 8$ for disorder. This Z-score threshold provides the optimal agreement for a binary classification of order/disorder for all prediction methods on average (see Supplementary Fig. S10). A good predictor should optimize the fraction of correctly identified disordered residues (true positives, TP) while simultaneously minimizing the fraction of false positives (FP). ROC curves display TP vs. FP as a function of the probability threshold and the corresponding area under this curve (*AUC*) is an aggregate measure of the quality of a predictor that is not affected by any skew/bias of the estimated probabilities. A perfect classifier would yield $AUC = 1$, whereas random guessing gives $AUC = 0.5$. ROC curves for all predictors are shown in Fig. 3 and the *AUC* values are listed in Table 1. The non-binary methods display *AUCs* ranging from 0.733 (MetaDisorder3D) to 0.890 (SPOT-disorder) and reiterate the trends described above for the ranking of the predictors (see Table S1 and Supplementary Fig. S7).

It is apparent from Fig. 2 that some predictors are continuous, while other are more bimodal. In addition, for some methods predictions cluster on one side, suggesting a prediction bias. To quantify this bias of over-predicting order or disorder, the average probability of predicting low Z-scores (pZL for Z-scores < 8.0) and high Z-scores (pZH, Z-scores > 8.0) was calculated for each method. An unbiased method would have an average probability pZA = (pZL + pZH)/2 close to 0.5. At the same time, methods with good discrimination between order and disorder will display a large probability difference, pZD = pZL − pZH. Figure 4 plots the average probability (pZA, bias) against the probability difference. It is seen that DISOPRED2[4], DisEMBL

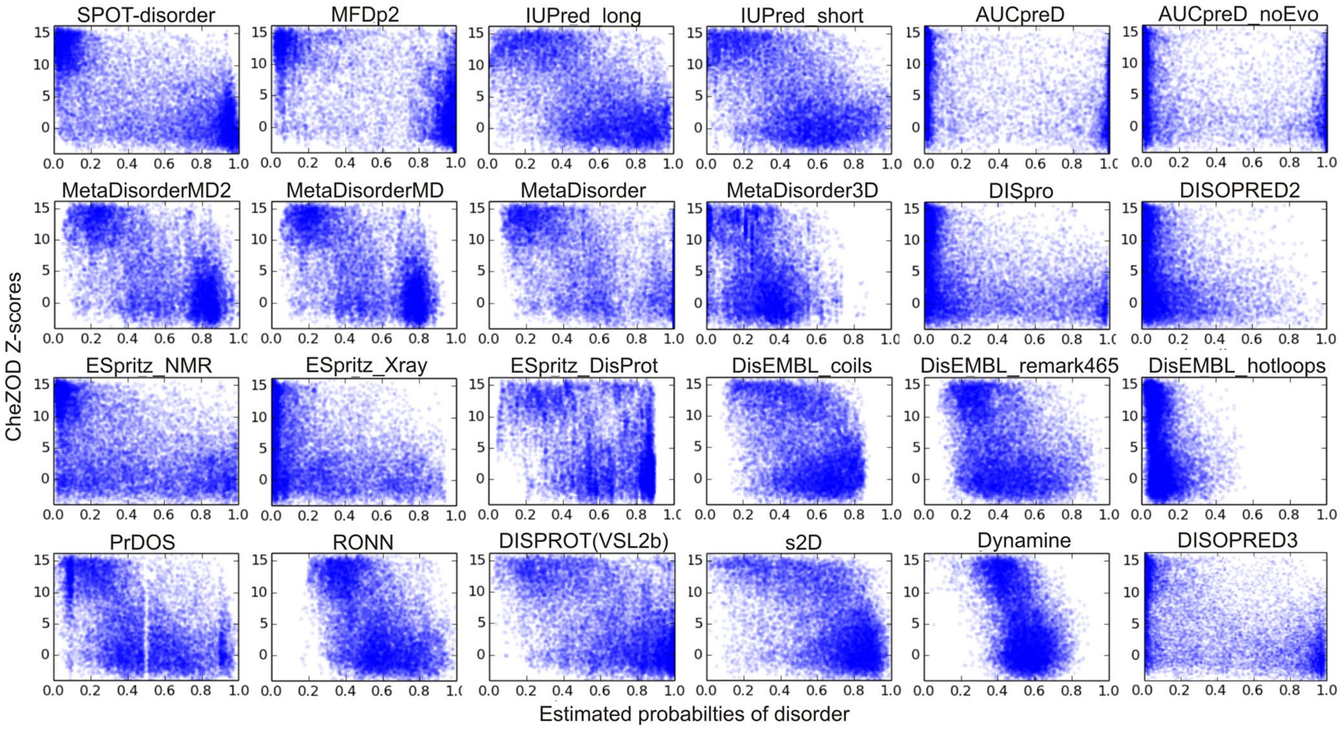

Figure 2. Z-score vs. estimated probability of disorder (p) for 24 continuous-valued prediction methods.

hotloops[61] and GlobPlot[62] are biased towards under-predicting disorder (e.g. using pZA < 0.3). On the other hand, no methods over-predict disorder (no method has pZA > 0.7). Along the other axis, SPOT-disorder has the highest probability difference suggesting the best (formal) discrimination between order and disorder. The above findings are mirrored in a classical confusion-based analysis (see Supplementary Table S2) except that for DISOPRED2 and DisEMBL hotloops a probability cut-off different from p = 0.5 was used, and therefore no significant over-prediction of order was found by this analysis. GlobPlot and ESpritz-Xray[57] methods have False Negative Rates (FNRs) as high as 0.98 and 0.718, respectively, but at the other end of the extreme, the methods with the highest False Positive Rates (FPRs), ESpritz_DisProt and DISPROT[63] (VSL2b), have FPRs of 0.415 and 0.401, respectively, and do not over-predict disorder to a similar extent.

Discussion

Residues with missing X-ray densities are relatively rare, with only 2.4% of the residues being non-observed in the dataset tested here (see Methods) and 8.6% in a set used for training SPOT-disorder[55] (See Supplementary Discussion and Supplementary Table S2) and the disordered regions identified in X-ray data are relatively short (Supp.Table S2). Conversely, long regions of disordered residues as well as completely disordered proteins are abundant in the DisProt database[19,20] (see Supplementary Discussion). This pronounced difference between the two data sources has long been realized, and complementary methods dedicated to predicting either short or long regions of disorder have been developed by training on X-ray or DisProt data, respectively[57,64,65]. Interestingly, yet maybe not surprising, dedicated subversions of predictors show the best performance when evaluated on the same type of data as were used for training[21,66]. To elaborate on this, the CheZOD database was divided into different subsets chosen as to represent data sets with different characteristics as e.g. content of disorder and size of disordered regions (see Supplementary Discussion). It was found that the ranking of the prediction methods was generally preserved and that the performance on the different subsets reflect the data used for training the methods (see Supplementary Discussion and Supplementary Figs S8 and S9). Since the CheZOD database is diverse and balanced, containing both structured proteins with short and long disordered loops as well as completely disordered proteins[22], it is ideal for assessing the performance of predictors of general disorder of no particular flavor.

NMR-derived Z-scores for proteins in the CheZOD database have been applied here in an attempt to rigorously benchmark the performance of a large number of disorder predictors (see Table 1). Contrary to CASP chronological extrapolations outlined above, it was found that the most recent predictors feature improved performance. Notably, the newer implementation of DISOPRED, DISOPRED3[67], performs significantly better than the older version, DISOPRED2[4]. Several trends in the performance of the predictors related to the type of inputs and optimization procedure were observed. Older methods and methods that focus on speed use only amino acid (AA) sequence-based features, such as AA composition, physiochemical properties, interaction energies and sequence complexity, and display comparatively less good performance. Inclusion of evolutionary information derived from multiple sequence alignment profiles expands the repertoire with complementary features. The group of predictors here that use evolutionary (Evo) information generally perform better than the predictors

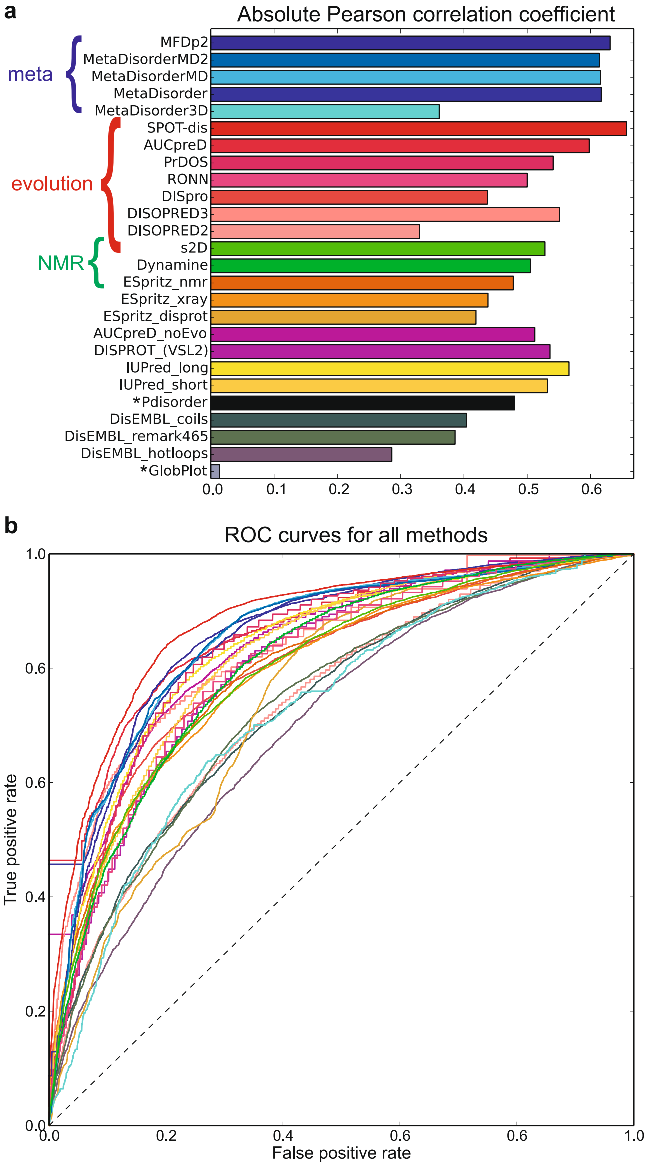

Figure 3. (**a**) Ranking of disorder prediction methods according to the absolute Pearson linear correlation coefficient between estimated disorder probability and Z-score shown as a histogram. The order of the methods is as in Table 1 (see Supplementary Fig. S6 for Spearman correlation). Annotation with colored curly brackets highlight meta-methods (meta), methods that apply information from evolutionary profiles of aligned sequences (evolution), and methods that use NMR data for training (NMR). Asterisks mark the binary prediction methods. (**b**) Receiver-Operating Characteristics (ROC) curves for all non-binary predictors for using estimated disorder probability to predict Z-score under/above the threshold Z = 8. Colors as in the histogram. The corresponding area under the curve (AUCs) are provided in Table 1 and shown as histograms in Supplementary Fig. S7. Note that the edgy appearance of some of the ROC curves are due to fewer decimal points on the estimated probabilities of disorder.

without it (Table 1). Finally, the metapredictors that use estimated disorder probabilities from other predictors display very good performance.

To compare the authority of different data-sources to judge disorder, we perform a comparison across data for the same methods by deriving traditional binary classifier metrics; the AUC and the Mathews correlation coefficient (MCC) (see e.g.[18]). MCC is a balanced measure of correlation that considers false and true positives as well as their negatives. The AUC and MCCs were calculated and compared to values reported in the literature for testing against DisProt[21] and X-ray data, as summarized previously[9] (see Table S1 and Fig. 5). We find that values for both AUC and MCC are significantly higher for the same predictors when compared to the DisProt and X-ray evaluation sets, respectively (Fig. 5 and Table S1). This strongly suggests that the CheZOD Z-score classifier is more predictable and more accurate, in the sense that it contains fewer miss-classifications.

The analysis presented here provides a guideline for selecting the most appropriate predictor for assessing disorder and to avoid intrinsic bias. As a point in case, DISOPRED2 was used to estimate the content of disordered

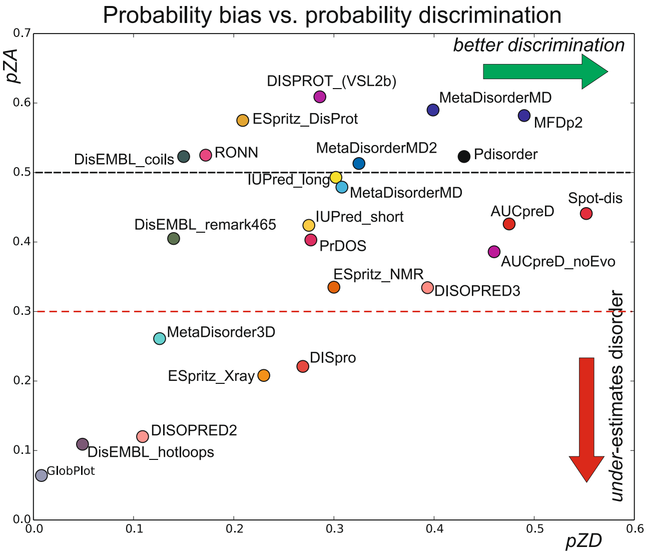

Figure 4. Probability bias vs. probability discrimination showing pZA as a function of pZD (see text and Online Methods). Each predictor is shown with a circle using the same colors as in Fig. 3 above. A black broken line corresponding to a completely unbiased predictor with pZA = 0.5 is shown for reference. Predictors below the red dashed line (pZA = 0.3) considerably under-estimate disorder. Methods that noticeably over-predict disorder (i.e. pZA > 0.7) were not observed.

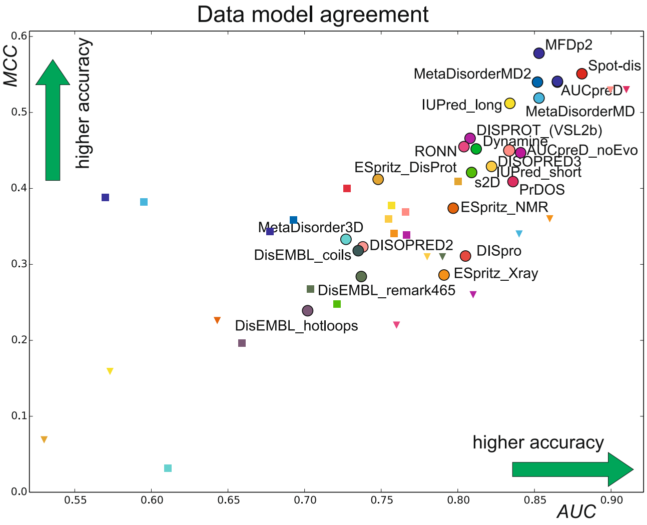

Figure 5. Performance of predictors on different data sets. MCC vs. AUC is shown for each method tested here with a circle using same colors as in Fig. 3 and compared to values reported in the literature for methods also tested here depicted as squares when tested against DisProt data and triangles when tested against X-ray data. See Supplementary Table S1 for all numbers and details related to DisProt and X-ray source data (Note that some of the prediction methods analyzed here were not included in the corresponding studies tested against DisProt or X-ray data, and hence, there are fewer labels of triangles and squares).

residues in various proteomes revealing a content of ca. 33% in Eukaryotes[4]. Importantly, our analysis now shows that DISOPRED2 markedly *under*-predicts disorder, suggesting that protein disorder in eukaryotes is even more prevalent than previously assumed.

Conclusions

We have demonstrated that validated, balanced NMR chemical shift data of proteins can be used to benchmark widely-used disorder predictors. Cross-data comparison of the performance for the same predictors demonstrated that the CheZOD dataset is more appropriate than previously utilized sources. A detailed analysis revealed that the most recent and most advanced prediction methods display the best performance, and bias for under-predicting disorder was evaluated quantitatively. We provided several performance measures to help researchers make an informed decision for selection of the most appropriate disorder prediction method.

Methods

Production of disorder probabilities for the proteins in the benchmarking set. The genesilico metaserver (http://iimcb.genesilico.pl/metadisorder/) was used to obtain estimated probabilities of disorder for a range of different disorder prediction methods (see Table 1 in main text) including their own meta-predictors. Furthermore, we added predictions from several other methods where parallel batch job submission was possible using their servers: SPOT-disorder[55], MFDp2[14], AUCpreD[56], three versions of ESpritz[57] based on different training data, *viz*. X-ray missing density, NMR ensemble structural disorder classification and DisProt disorder. Classic DisEMBL binary predictions were replaced by continuous predictions using the automatic job submission system at http://dis.embl.de. DynaMine[59] and s2D[58], which predict continuous NMR data, were also included. Predictions of populations of secondary structure types from *s2D* were interpreted using the sum of the estimated populations of alpha-helix and beta-sheet as a probability of order, as before[24]. Predictions of the order parameter S^2 from *Dynamine* were converted to a probability of disorder using the bijective transformation $p = \sqrt{1 - S^2}$. To summarize, the prediction methods tested were: MetaDisorder including MD/MD2/3D variants[13], SPOT-disorder[55], AUCpreD (with/without evolution)[56], MFDp2[14], PrDOS[68], RONN[69], DISpro[70], DISOPRED2[4], DISOPRED3[67], *s2D*[58], DynaMine[59], ESpritz NMR/Xray/DisProt variants[57], DISPROT[63] (VSL2b) (also referred to as PONDR), IUPred long/short variants[60], Pdisorder (http://www.softberry.com/), DisEMBL coils/remark465/hotloops variants[61] and GlobPlot[62].

The set of structured proteins with chemical shifts. The database of structured proteins described before[49] was used. However, in the present study we did not exclude entries homologous to proteins from the CheZOD database leading to a final set of 896 proteins with assigned chemical shifts. From this set, 222 proteins structures were determined by X-ray crystallography whereas the remaining 674 were determined by NMR spectroscopy. A trimmed unbiased set of X-ray structures was derived from the set of 222 proteins by removing entries if (i) the biologically significant oligomerization state was not a monomer, (ii) larger ligands were present, (iii) the protein sequence of the X-ray structure and the corresponding sequence of assigned chemical shifts differed for more than 10% of the residues. These criteria resulted in a reduced database of 90 entries. For both sets of X-ray structures, residues in the X-ray sequence (SEQRES record) that were absent in the coordinate section (i.e. those mentioned in the REMARK 465) were identified. Following this procedure, we identified 717 missing residues in the set of 222 X-ray structures compared to 30495 residues that were observed in the structure - and similarly 234/13581 for the reduced 90 entries set. Note that only residues with assigned chemical shifts in the corresponding NMR study were included in the above analysis. Within the set of entries corresponding to NMR structures, the 100 with the highest fraction of residues with CheZOD Z-scores < 5.0 were selected and used for comparison with the parameters (see below) describing structural variation in the corresponding NMR ensemble of structures. Furthermore, we identified 23 proteins from the refDB database[71] described above having chemical shifts assigned for all backbone atom types that had available simulated molecular dynamics trajectories in the Dynameomics database[53,54]. The Z-scores were compared to the rms Cα coordinate fluctuations within the MD trajectories for these proteins.

Definition of torsion angle and coordinate variations and order parameters. The dihedral angle order parameter S_{HW} of Hyberts, Wagner and co-workers[72] is defined as:

$$S_{HW}(\theta) = \frac{1}{N}\sqrt{\left(\sum_{i=1}^{N}\sin(\theta_i)\right)^2 + \left(\sum_{i=1}^{N}\cos(\theta_i)\right)^2} \tag{1}$$

for an ensemble of N structures, where θ_i is the value of a particular dihedral angle θ in the i^{th} member of the ensemble. Based on the backbone dihedral angles ϕ and ψ, the sequence-specific backbone dihedral angle parameter, D_i, for residue i in a protein sequence is defined as:

$$D_i = \frac{1}{6}\sum_{j=i-1,\ i,i+1}(S_{HW}(\phi_j) + S_{HW}(\psi_j)) \tag{2}$$

This order parameter is converted to a torsion angle standard deviation, s(i), using the approximate relation[72]:

$$s(i) = 2\arccos\left(1 + \frac{\ln(D_i)}{2}\right) \tag{3}$$

A parameter describing the variation in Cartesian coordinates for a specific residue is derived from the inter-atomic variance matrix (IVM) following a procedure akin to the FindCore algorithm[73]. Each element, v_{ij} in the variance matrix is defined as:

$$v_{ij} = \frac{1}{N}\sum_{k=1}^{N}\left(d_{ijk} - \bar{d}_{ij}\right)^2,\ \ \bar{d}_{ij} = \frac{1}{N}\sum_{k=1}^{N}d_{ijk} \tag{4}$$

where d_{ijk} is the Cα(i)-Cα(j) distance for conformer, k, in the ensemble.

Each row, v_i, in the matrix, excluding diagonal and next-to-diagonal elements, v_{ii} and v_{ij} with |i-j| = 1 is sorted numerically and indexed by increasing rank:

$$\lambda_{i1} < \lambda_{i2} < \cdots < \lambda_{in} \tag{5}$$

where $\lambda_{ij} = v_{iq}$ is j'th smallest element of the row v_i and n denotes the total number of such variance elements – i.e. the number of residues minus 3.

The residue coordinate variation, $t(i)$, is then calculated as the weighted average:

$$t(i) = \frac{\sum_{j=1}^{n} w_j \sqrt{\lambda_{ij}}}{\sum_{j=1}^{n} w_j}, \quad w_j = e^{-\beta\left(\frac{j}{n}\right)^2} \tag{6}$$

where $\beta = 10.0$ is used here. The parameters, s and t, describing the residue angle and coordinate variation, respectively, are then converted to the corresponding order parameters, S and T, using:

$$\mathrm{S} = \frac{1}{(1 + (\frac{\mathrm{s}}{\mathrm{s}_0})^2)} \text{ and } \mathrm{T} = \frac{1}{(1 + (\frac{\mathrm{t}}{\mathrm{t}_0})^2)} \tag{7}$$

where $s_0 = 75°$ and $t_0 = 1.5$ Å were used here as the reference values.

The Jensen-Shannon divergence, JSD[74], describes the similarity between two (discrete) probability distributions, P and Q.

$$JSD(P, Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{8}$$

where M is the average of the distributions

$$M = \frac{1}{2}(P + Q) \tag{9}$$

and D is the Kullbeck-Leibner divergence:

$$D(P||M) = \sum_i P(i)\log\left(\frac{P(i)}{M(i)}\right) \tag{10}$$

here we calculate JSD for the distributions of Z-scores corresponding to above/below reference values $s_0 = 1.5$ Å and $t_0 = 75°$ for the residue angle and coordinate variation, respectively, and for residues corresponding to observed residues in X-ray structures vs. missing residues (REMARK 465).

Data Availability

The full database containing protein sequences, BMRB id, and CheZOD Z-scores is available at http://www.protein-nmr.org./.

References

1. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197–208 (2005).
2. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**, 18–29 (2015).
3. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589–6631 (2014).
4. 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).
5. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**, 215–246 (2008).
6. Romero, P., Obradovic, Z. & Dunker, A. K. Natively disordered proteins: functions and predictions. *Appl Bioinformatics* **3**, 105–113 (2004).
7. Midic, U., Oldfield, C. J., Dunker, A. K., Obradovic, Z. & Uversky, V. N. Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome. *Protein Pept Lett* **16**, 1533–1547 (2009).
8. 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).
9. Meng, F. C., Uversky, V. N. & Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cellular and Molecular Life Sciences* **74**, 3069–3090 (2017).
10. Oates, M. E. *et al.* D^{2}P^2: database of disordered protein predictions. *Nucleic Acids Research* **41**, D508–D516 (2013).
11. Piovesan, D. *et al.* MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Research* **46**, D471–D476 (2018).
12. Di Domenico, T., Walsh, I. & Tosatto, S. C. E. Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. *Bmc Bioinformatics* **14** (2013).
13. Kozlowski, L. P. & Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* **13**, 111 (2012).
14. Mizianty, M. J., Peng, Z. & Kurgan, L. MFDp2. *Intrinsically Disordered*. *Proteins* **1**, e24428 (2013).
15. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L. & Rost, B. Improved Disorder Prediction by Combination of Orthogonal Approaches. *Plos One* **4** (2009).
16. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–v (1995).
17. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A. & Kryshtafovych, A. Evaluation of disorder predictions in CASP9. *Proteins* **79**(Suppl 10), 107–118 (2011).
18. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A. & Fidelis, K. Assessment of protein disorder region predictions in CASP10. *Proteins* **82**, 127–137 (2014).
19. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* **35**, D786–793 (2007).
20. Piovesan, D. *et al.* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* **45**, D219–D227 (2017).
21. 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).

22. Nielsen, J. T. & Mulder, F. A. A. There is Diversity in Disorder—"In all Chaos there is a Cosmos, in all Disorder a Secret Order". *Frontiers in Molecular Biosciences* **3** (2016).
23. Toth-Petroczy, A. *et al*. Structured States of Disordered Proteins from Genomic Sequences. *Cell* **167**, 158–170.e112 (2016).
24. Sormanni, P. *et al*. Simultaneous quantification of protein order and disorder. *Nat Chem Biol* **13**, 339–342 (2017).
25. Wuthrich, K. Protein-structure determination in solution by nmr-spectroscopy. *J Biol Chem* **265**, 22059–22062 (1990).
26. Wagner, G., Hyberts, S. G. & Havel, T. F. NMR structure determination in solution - a critique and comparison with x-ray crystallography. *Ann Rev Biophys Biomol Struct* **21**, 167–198 (1992).
27. Brunger, A. T. & Nilges, M. Computational challenges for macromolecular structure determination by x-ray crystallography and solution nmr-spectroscopy. *Q Rev Biophys* **26**, 49–125 (1993).
28. Guntert, P. Structure calculation of biological macromolecules from NMR data. *Q Rev Biophys* **31**, 145–237 (1998).
29. Wuthrich, K. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angew Chem Int Ed* **42**, 3340–3363 (2003).
30. Palmer, A. G., Kroenke, C. D. & Loria, J. P. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Nucl Magn Reson. Biol Macromol, Pt B* **339**, 204–238 (2001).
31. Palmer, A. G. NMR characterization of the dynamics of biomacromolecules. *Chem Rev* **104**, 3623–3640 (2004).
32. Mittermaier, A. & Kay, L. E. Review - New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228 (2006).
33. Ulrich, E. L. *et al*. BioMagResBank. *Nucleic Acids Research* **36**, D402–D408 (2008).
34. Felli, I. C. & Pierattelli, R. Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* **64**, 473–481 (2012).
35. Brutscher, B. *et al*. NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines. *Adv Exp Med Biol* **870**, 49–122 (2015).
36. Wishart, D. S. & Sykes, B. D. Chemical-shifts as a tool for structure determination. *Nucl Magn Reson, Pt C* **239**, 363–392 (1994).
37. Wishart, D. S. & Case, D. A. Use of chemical shifts in macromolecular structure determination. *Nucl Magn Reson. Biol Macromol, Pt A* **338**, 3–34 (2001).
38. Berjanskii, M. V. & Wishart, D. S. A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts. *J Ame Chem Soc* **127**, 14970–14971 (2005).
39. Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci* **15**, 2795–2804 (2006).
40. Camilloni, C., De Simone, A., Vranken, W. F. & Vendruscolo, M. Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry* **51**, 2224–2231 (2012).
41. Kjaergaard, M. & Poulsen, F. M. Disordered proteins studied by chemical shifts. *Prog Nucl Magn Reson Spectrosc* **60**, 42–51 (2012).
42. Tamiola, K. & Mulder, F. A. A. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem Soc Trans* **40**, 1014–1020 (2012).
43. Kragelj, J., Ozenne, V., Blackledge, M. & Jensen, M. R. Conformational propensities of intrinsically disordered proteins from NMR chemical shifts. *Chemphyschem* **14**, 3034–3045 (2013).
44. Best, R. B. & Lindorff-Larsen, K. Editorial overview: Theory and simulation: Interpreting experimental data at the molecular level. *Curr Opin Struct Biol* **49**, IV–VI (2018).
45. Showalter, S. A. & Bruschweiler, R. Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *J Chem Theo Comput* **3**, 961–975 (2007).
46. Joerger, A. C. & Fersht, A. R. In *Annu Rev Biochem* Vol. 77 *Annu Rev Biochem* 557–582 (2008).
47. 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** (2008).
48. Meek, D. W. Regulation of the p53 response and its relationship to cancer. *Biochem J* **469**, 325–346 (2015).
49. Nielsen, J. T. & Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J Biomol NMR* **70**, 141–165 (2018).
50. Uversky, V. N. p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. *Int J Molec Sci* **17** (2016).
51. Xue, B., Brown, C. J., Dunker, A. K. & Uversky, V. N. Intrinsically disordered regions of p53 family are highly diversified in evolution. *Biochim Biophys Acta - Proteins and Proteomics* **1834**, 725–738 (2013).
52. Ayed, A. *et al*. Latent and active p53 are identical in conformation. *Nat Struct Biol* **8**, 756–760 (2001).
53. Benson, N. C. & Daggett, V. Dynameomics: Large-scale assessment of native protein flexibility. *Protein Sci* **17**, 2038–2050 (2008).
54. van der Kamp, M. W. *et al*. Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure* **18**, 423–435 (2010).
55. 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).
56. Wang, S., Ma, J. & Xu, J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* **32**, i672–i679 (2016).
57. Walsh, I., Martin, A. J., Di Domenico, T. & Tosatto, S. C. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
58. Sormanni, P., Camilloni, C., Fariselli, P. & Vendruscolo, M. The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. *J Mol Biol* **427**, 982–996 (2015).
59. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. & Vranken, W. F. From protein sequence to dynamics and disorder with DynaMine. *Nat Commun* **4**, 2741 (2013).
60. 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).
61. Linding, R. *et al*. Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
62. 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).
63. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573–584 (2003).
64. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y. & Noguchi, T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046–2053 (2007).
65. Shimizu, K., Hirose, S. & Noguchi, T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics* **23**, 2337–2338 (2007).
66. Walsh, I. *et al*. Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* **31**, 201–208 (2015).
67. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
68. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* **35**, W460–464 (2007).
69. Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376 (2005).

70. Cheng, J., Sweredoski, M. J. & Baldi, P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Min. Knowl. Discov.* **11**, 213–222 (2005).
71. Zhang, H. Y., Neal, S. & Wishart, D. S. RefDB: A database of uniformly referenced protein chemical shifts. *J Biomol NMR* **25**, 173–195 (2003).
72. Hyberts, S. G., Goldberg, M. S., Havel, T. F. & Wagner, G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Sci* **1**, 736–751 (1992).
73. Snyder, D. A. & Montelione, G. T. Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins: Structure, Function, and Bioinformatics* **59**, 673–686 (2005).
74. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans Inform Theory* **37**, 145–151 (1991).
75. Hrabe, T. *et al*. PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res* **44**, D423–D428 (2016).
76. Canadillas, J. M. *et al*. Solution structure of p53 core domain: structural basis for its instability. *Proc Natl Acad Sci USA* **103**, 2109–2114 (2006).
77. Rowell, J. P., Simpson, K. L., Stott, K., Watson, M. & Thomas, J. O. HMGB1-facilitated p53 DNA binding occurs via HMG-Box/p53 transactivation domain interaction, regulated by the acidic tail. *Structure* **20**, 2014–2024 (2012).
78. Wong, T. S. *et al*. Biophysical characterizations of human mitochondrial transcription factor A and its binding to tumor suppressor p53. *Nucleic Acids Res* **37**, 6765–6783 (2009).

Acknowledgements

We thank Dr. Lukasz P. Kozlowski for help with batch submissions on the genesilico server.

Author Contributions

The project was designed and developed by F.A.A.M. and J.T.N. J.T.N. performed the mathematical and statistical analysis of the data and produced the figures. The paper was written by F.A.A.M. and J.T.N.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41644-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.