



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations


Graduate School

2021

Computational Analysis and Prediction of Intrinsic Disorder and Intrinsic Disorder Functions in Proteins

Akila I. Katuwawala
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Biochemistry Commons](#), [Bioinformatics Commons](#), [Computational Engineering Commons](#), and the [Structural Biology Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/6645>

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

©Akila Iimesha Katuwawala, May 2021

All Rights Reserved.

Computational Analysis and Prediction of Intrinsic Disorder and Intrinsic Disorder Functions in Proteins

A dissertation submitted in partial fulfillment of the requirements for the degree of Engineering, Doctor of Philosophy with a concentration in Computer Science at Virginia Commonwealth University

by

AKILA IMESHA KATUWAWALA

Master of Science, University of Westminster, United Kingdom, 2017

Bachelor of Science, University of Colombo, Sri Lanka, 2017

Bachelor of Science, University of Greenwich, United Kingdom, 2015

Director: Lukasz Kurgan,

Professor, Department of Computer Science

Virginia Commonwealth University

Richmond, Virginia

May, 2021

Acknowledgements

I would like to thank my advisor Dr. Lukasz Kurgan for his sincere guidance and support during past few years to make my PhD journey a memorable and successful experience. I would like to express my gratitude to Dr. Vladimir Uversky, Dr. FangXiang Wu, Dr. Tomasz Arodz and Dr. Bartosz Krawczyk for serving in my dissertation committee and for their valuable feedback.

I would also like to thank all the collaborators from other institutions and my colleagues from Kurgan lab for their support. I would like to acknowledge the financial assistance provided by the National Science Foundation during my PhD studies. Last but not least I would like to thank my family and friends for their love and encouragement.

Table of Contents

Acknowledgements.....	i
Table of Contents.....	ii
List of Tables	vi
List of Figures	ix
List of Acronyms.....	xvi
Abstract.....	xvii
Chapter 1. Introduction	1
1.1 Motivation.....	2
1.2 Objectives and Proposed Contributions	3
1.3 Organization of the dissertation	4
Chapter 2. Background and Related Work.....	7
2.1 Introduction to proteins, intrinsic disorder and its functions.....	7
2.1.1 Intrinsic disorder in proteins.....	7
2.1.2 Functional aspect of intrinsically disordered proteins	9
2.2 Residue level protein structure and function prediction in disorder proteins.....	12
2.2.1 Computational prediction of intrinsic disorder in proteins.....	12
2.2.2 Computational prediction of functions in intrinsically disordered proteins	17
2.2.3 Computational modelling related to the development of disorder predictors	21
2.2.4 Evaluation of predictive performance	22
Chapter 3. Elucidation and comparative analysis of protein-level predictive performance for current disorder predictors.....	25
3.1 Related work in disorder predictor assessment	25

3.2	Benchmark datasets and selection of disorder predictors	28
3.3	Predictive performance of disorder predictors	31
3.3.1	Dataset level predictive performance	31
3.3.2	Protein level predictive performance	31
3.4	Complementarity and relative protein-level performance of disorder predictors	36
3.5	Case Study	40
3.6	Summary	41
Chapter 4. Development of a novel protein-level predictor recommendation system to improve predictive performance of disorder predictions		
4.1	Protein-level predictive performance of disorder predictors.....	44
4.2	Experimental workflow	46
4.2.1	Architecture and design of the protein-level predictors of disorder prediction quality	46
4.2.2	Design of the protein-level disorder predictor recommendation system	50
4.2.3	Analysis of the predictive model	50
4.3	Assessment of the protein-level disorder predictor recommendation system	53
4.3.1	Predictive performance of the extra tree regressor models.....	53
4.3.2	Use of the extra tree regressors for the selection of well-predicted proteins.....	55
4.3.3	Predictive quality of DISOselect.....	59
4.4	DISOselect webserver	62
4.5	Summary	63
Chapter 5. Assessment and comparative analysis of the predictive performance of disorder predictions for specific functional types of disordered proteins		
5.1	Overview of the past intrinsic disorder predictor assessments	65
5.2	Selection of disorder predictors.....	69

5.3	Collection of benchmark dataset	70
5.4	Comparative assessment of predictive performance	71
5.4.1	Effect of the sequence similarity reduction and structured region validation on benchmark dataset.....	71
5.4.2	Comparative assessment of disorder predictors on the benchmark dataset	75
5.4.3	Predictive performance assessment on the disordered protein-binding and nucleic acid-binding proteins.....	76
5.5	Summary	78
Chapter 6. Accurate prediction of the disordered lipid-binding residues from protein sequences		81
6.1	Introduction and motivation	81
6.2	Materials and methods	83
6.2.1	Dataset description.....	83
6.2.2	DisoLipPred architecture	85
6.3	Results	90
6.3.1	Ablation analysis of the prediction model.....	90
6.3.2	Comparative assessment on the test dataset	92
6.3.3	DisoLipPred predictions on the <i>Saccharomyces cerevisiae</i> proteome.....	95
6.3.4	DisoLipPred prediction assessment on transmembrane proteins	97
6.3.5	Case study	98
6.3.6	Webserver.....	99
6.3.7	Summary	100
Chapter 7. Summary		101
7.1	Major contributions	103
7.2	List of related publications.....	105

References	107
Glossary.....	119
Appendix 1 – Complete set of 130 features used to implement the disorder predictor recommendation system (the DISOselect method)	120
Appendix 2 –DisoLipPred supplementary data	125

List of Tables

Table 1: Number of the molecular function annotations for the functionally annotated IDRs in the DisProt database. Annotations tagged as ambiguous have been excluded.	11
Table 2: Number of molecular partner annotations for the functionally annotated IDRs in the DisProt database.	12
Table 3: Summary of the nine highly-cited computational disorder predictors. The number of citations was collected from the Web of Science as of May 2020. The methods are sorted in the ascending chronological order.	16
Table 4: Classification, citations and availability of the current predictors of IDR functions. The methods are <i>categorized</i> based on their predictive target (molecular partner vs. molecular function) and sub-type of the target (protein, DNA and RNA for molecular partners vs. flexible linker and moonlighting region for molecular functions). Predictors are sorted within each sub-type by the year of publication. The citations and availability are based on information as of Feb 25, 2019. The citations were collected using Google Scholar, where the annual citations are computed as an average number of citations per year since publication. Methods without any availability are listed as “not available” and those for which the websites cannot be found are denoted as “no longer available”.	19
Table 5: Dataset- and protein-level predictive quality for the 13 considered disorder predictors. The dataset-level accuracy and AUC are compared against the previously published results. We note that false positive rates are typically not reported in the past studies. Protein-level results are summarized with the median value. The methods are sorted by their dataset-level AUC on our benchmark dataset. These results were published in [124].	30
Table 6: Predictive performance of the extra tree regressor-based model and two controls. Mean squared error (MSE) and Pearson correlation coefficients (PCC) values are calculated between the predicted AUC and the actual AUC for each test proteins. Controls were produced using a random and a sequence similarity-based approaches. Paired significance tests were	

performed between the predicted AUCs of our regressor and the results produced by the controls: [+] denotes that our model is significantly better with p -value $<.05$. We used the paired t test for normal measurements and the Wilcoxon test otherwise; normality was tested with the Anderson-Darling test at the .05 significance. These results were published in [139]. 54

Table 7: Comparison of the per-protein AUC values produced by the 12 disorder predictors, the oracle method that selects the predictor with the highest AUC and the selection based on the highest putative AUC produced by DISOselect for the test proteins. We compared the mean per-protein AUCs computed over the test proteins and the AUCs for the worst (the least accurately predicted) quartile of the test proteins (i.e., the 25% point in Figure 17). Methods are sorted by their mean per-protein AUCs. Significance of the differences in the per-protein AUCs of the predictions selected by DISOselect and the predictions generated by the other methods (including the oracle) was assessed with the t -test for normal measurements and the Wilcoxon test otherwise; normality was tested with the Anderson-Darling test at 0.05 significance; we sampled 50% of proteins in the test dataset ten times at random and compared the corresponding 10 pairs of AUCs; the resulting p -values are listed in the last column. These results were published in [139]. 61

Table 8: Summary of the past comparative assessments of disorder predictors. The articles are sorted chronologically (from the most recent). The citation numbers were collected from Google Scholar on September 29, 2020. Predictors shown in the bold font in the “suggested best disorder predictors” column are included in the comparative assessment in this section. 67

Table 9: Summary of the benchmark dataset. 71

Table 10: Predictive performance on the new benchmark dataset. The table lists results on the complete benchmark dataset with 357 proteins, the set of 38 fully disordered proteins, the set of 38 fully structured proteins, and the benchmark dataset of 319 proteins that exclude the fully structured proteins. We quantify statistical significance of differences in AUC between the best predictor (identified in bold font) and each the other nine predictors on a given dataset. We bootstrap 50% of the proteins 100 times. For normal measurements (tested with the Anderson-Darling test at 0.05 significance) we use the paired t -test; otherwise we use the

Wilcoxon rank sum test; = and + mean that the differences are not significant (p-value > 0.01) and significant (p-value ≤ 0.01), respectively. 74

Table 11: Description of the training and test datasets. 84

Table 12: Experimental setups for the ablation study. 91

Table 13: Predictive performance of DisoLipPred and its variants from the ablation analysis (Table 1) on the test dataset. We perform the assessment on the complete test dataset, and also on the subset of disordered residues from the test dataset. We quantify the binary metrics (sensitivity and F1) at the fixed specificity = 0.9. We assess the statistical significance of the differences between the results produced by DisoLipPred and each of the variants using procedure explained in Section 2.2. * indicates that DisoLipPred provides significantly better result (p-value <0.05). 91

Table 14: Predictive performance on the test dataset. We perform the assessment on the complete test dataset, and also on the subset of the native disordered residues from the test dataset. We quantify the binary metrics (sensitivity and F1) at the fixed specificity = 0.9. for the predictors that produce the propensity scores. We use the default sensitivity, F1, and specificity values for the other three methods that produce only binary predictions: SCAMPI 2, Phobius and BLAST. We assess the statistical significance of the differences between the results produced by DisoLipPred and every other tool using procedure explained in Section 2.2. * indicates that DisoLipPred provides significantly better result (p-value <0.05). Methods are sorted in the ascending order by their AUC within each predictor type group. 93

Table 15: Predictive performance on the TM dataset. The performance is measured assuming that the native transmembrane regions constitute positive annotations. Both transmembrane predictors (SCAMPI 2 and Phobius) produce only binary predictions and thus their prediction rate cannot be calibrated. Instead, we calibrate the rate of the DisoLipPred’s predictions to match the specificity of SCAMPI 2 and Phobius. 98

List of Figures

- Figure 1:** The PyMOL visualization of NMR structure of human Bcl-2-like protein 1 (PDB ID: 2ME8). All 20 different models are overlapped, and single color denotes a single model..... 9
- Figure 2:** Panel A is three-dimensional structure of the human Bcl-2-like protein 1. The red and blue regions denote native (experimentally annotated) ordered and disordered regions, respectively. Panel B visualizes disorder and disorder function prediction from IUPred-Long, IUPred-Short, SPOT Disorder and DFLpred. The top horizontal bar gives the actual/native disorder and order annotations while the panels below show the putative propensities and the corresponding binary predictions (gray and pink horizontal bars) generated by the four predictors. The presence of a bar denotes presence of putative disorder or specific disorder function in the corresponding region. 13
- Figure 3:** Comparison of the benchmark (dataset-level) and protein-level predictive performance measured with AUC for three disorder predictors (GlobPlot, IUPred and VSL2B) and five color-coded proteins: non-toxic nonhemagglutinin type D (green marker; UniProt Q9LBR2), guanylate kinase (red marker; UniProt Q2GLF7), alkaline phosphatase (violet marker; UniProt A1YYW7), hydroxymethyltransferase (orange marker; UniProt P0A5Q8) and phytoene desaturase (blue marker; UniProt P21685). The benchmark AUC values are shown on the x-axis, while the protein-level AUC values are color-coded, shown on the y-axis and their values are given next to the corresponding markers. The black isometric AUC line shows equivalent dataset-level and protein-level values. Published in [124] 27
- Figure 4:** Distributions of the protein-level predictive quality measured with accuracy (green plots), AUC (blue plots) and false positive rate (FPR; red plots) for the 13 disorder predictors. The y-axis gives the fraction of the proteins in a given range of accuracy/AUC/FPR values. These results were published in [124]. 32
- Figure 5:** Distributions of the protein-level predictive quality measured with accuracy (panel A), AUC (panel B) and false positive rate (panel C) for the 13 disorder predictors. Box plots show the second quartile (in red), median (between red and green boxes) and third quartile (in green) for the distribution of the protein-level values. The whiskers denote the corresponding

10th and 90th percentiles. The black horizontal lines show the benchmark dataset-level performance. The predictors are sorted by their median values of the predictive performance. These results were published in [124]. 33

Figure 6: Analysis of the easy- and hard-to-predict proteins for the 13 disorder predictors. The easy proteins are predicted with higher-than-expected accuracy or AUC, i.e. their protein-level accuracy (AUC) > dataset-level accuracy (AUC). The hard proteins are predicted with relatively low accuracy or AUC, i.e. their protein-level accuracy (AUC) < (dataset-level accuracy (AUC) – average margin of difference between disorder predictors). Bars represent the fraction of the easy proteins (green bars) and the hard proteins (red bars) when predictive performance is quantified with AUC (dark shade) and accuracy (light shade). Predictors are sorted by fraction of the easy proteins quantified with AUC (dark green bars). These results were published in [124]. 34

Figure 7: Relation between the protein-level predictive performance and the native disorder content. Panel A shows medians of the average (over the 13 predictors) accuracy and AUC for proteins grouped by their native disorder content, defined as the fraction of disordered residues in the sequence. The whiskers give the 10th and 90th percentiles of these averages. Panel B gives the distribution of the disorder content for the easy and hard proteins that are in common across the 13 predictors. The box plots show the 2nd quartile, median (black horizontal line and 3rd quartile for the distribution of the protein-level disorder content values. The whiskers denote the corresponding 10th and 90th percentiles. These results were published in [124]. 35

Figure 8: Comparison of the protein-level predictive performance between the 13 disorder predictors. Panel A summarizes comparison of the AUC values (on green background) and accuracies (on blue background). Panel B considers the false positive rates (on red background). Statistical significance of the differences between all pairs of methods was assessed with the t-test for normal measures and otherwise with the Wilcoxon rank-sum test. Normality was tested with the Anderson–Darling test at 0.05 significance. We assume that the difference in predictive performance for a given pair of predictors is significant if the corresponding P-value is <0.01. Arrows point to the methods that secure significantly better predictive performance (P

< 0.01). The P-values are shown for the pairs of methods that are not significantly different. These results were published in [124]. 37

Figure 9: Pearson correlation coefficients (PCCs) between the protein-level predictive performances for each pair of the considered 13 disorder predictors. Panels A and B quantify the performance with accuracy and AUC, respectively. Both correlation matrices are symmetric. The sorting of the predictors differs between the two panel and was optimized to highlight clusters of highly correlated methods. Values of the PCC are color-coded where red denotes no correlation ($PCC < 0.3$), yellow denotes modest correlation ($0.3 \leq PCC \leq 0.66$) and green corresponds to high correlation ($PCC > 0.66$). These results were published in [124]. 38

Figure 10: Contributions of the 13 disorder predictors to the production of the highly accurate predictions. Panel A quantifies the fraction of proteins for which a given method generates the highest predictive performance compared to all other disorder predictors. Panel B show the fraction of proteins for which a given number of predictors offer highly accurate predictions, i.e. predictive performance that is higher than the expected performance of the best method (the dataset-level performance of the best method). The inner and outer rings show results when using accuracy and AUC, respectively. These results were published in [124]. 39

Figure 11: A case study that compares disorder predictions for the hydroxymethyltransferase protein from *M. tuberculosis* (Uniprot id: P9WIL7) that were generated by five methods: VSL2B (dark green), SPOT-DISORDER (magenta), DISOPRED3 (orange), GlobPlot (lime) and IUPred-long (gray). The putative propensities are shown using the solid, color-coded lines. The corresponding binary predictions are given using the color-coded horizontal bars at the bottom of the figure; thresholds that are used to convert the propensities into the binary predictions are visualized with the dashed horizontal lines in the top part of the figure. The red and blue horizontal bar denotes the native annotation of disordered and structured regions, respectively, which were annotated using crystal structure (PDB ID: 1OY0). These results were published in [124]. 41

Figure 12: Distribution of per-protein AUC values 12 computational disorder predictors over 3,126 proteins in the benchmark dataset. These results were published in [139]. 45

Figure 13: Architecture of the proposed recommendation system. Panel (a) gives flowchart of the proposed model while panel (b) shows the corresponding pseudocode. This illustration was published in [139]. 47

Figure 14: Key predictive features used to predict AUC of the 12 disorder predictors. The predictive performance of individual features is quantified with the Pearson correlation coefficients (PCC) between feature values (horizontal lines) and the prediction output (actual AUC) for each disorder predictor (vertical lines) that were quantified on the training dataset. Detailed explanation of features is available in Appendix 1. PCC values are color-coded where dark green is for $|PCC| \geq 0.3$, light green for $|PCC|$ between 0.15 and 0.30, white for $|PCC| < 0.15$, and grey with 'x' symbol indicate the a given feature is not included in the model for that predictor. The direction of arrows reveals the sign of PCC where upwards arrows denote positive correlation while downward arrows denote negative correlation. These results were published in [139]. 51

Figure 15: Importance the five feature categories for the predictive models designed for the 12 disorder predictors. We used a three-step process to derive the scores for each predictive model. First, the information gain of individual features was calculated from the extra-tree regressors. Second, features were divided into the five classes and the information gain of the features in the same category was summed up. Third, the summed values were dived by the sum of the information gain values of all features in the same model. The last step allow for directly comparison of relative contributions of each feature category. These results were published in [139]. 53

Figure 16: The dataset-level actual AUC values for subsets of the test proteins that are sorted based on their AUCs values estimated by the extra tree regressors. Individual panels correspond to different disorder predictors. Points in each panel correspond to AUCs of the subsets of test proteins for which the estimated AUCs are above a given percentile of all estimated AUCs, that is, the 20 mark on the x-axis corresponds to the 80% of the test proteins that have estimated AUCs that are above the 20th percentile of estimated AUCs generated by the extra tree regressors. The left-most point corresponds to the result on the complete test dataset while the right-most point corresponds to the 5% of test proteins with the highest estimated AUCs. The

line is the third-degree polynomial fit into the measured data. These results were published in [139]. 56

Figure 17: Improvements in the actual ROC-AUC, PR-AUC, sensitivity, MCC and accuracy values computed as the difference between the values for subsets of the top 25% (in green), 50% (in orange) and 75% (in blue) of the test proteins selected based on their AUCs values estimated by extra tree regressors and the whole dataset-level AUCs. Positive values of the improvement indicate that AUC for the subsets of the test proteins are higher than for the complete test dataset. The box plots represent the distribution of the improvements across the 12 disorder predictors where whiskers corresponding to the minimal and maximal improvements and boxes denote the first, second and third quartiles. These results were published in [139]. 58

Figure 18: Comparison of the per-protein AUC values between the 12 disorder predictors, the selection of the best disorder predictor using the highest putative AUCs generated by DISOselect (thick black line), and the oracle method (thick red line), and two conventional meta predictors (thick yellow and blue lines) on the test proteins. The oracle method selects the disorder predictor with the highest AUC among the 12 disorder predictors. Lines show the per-protein AUCs that are sorted in the ascending order for each of the considered methods. These results were published in [139]. 60

Figure 19: Evaluation of the differences in the protein-level area under the receiver operating characteristic curves (AUCs) for the same test proteins between the predictions selected with DISOselect and the average AUC of the 12 disorder predictors (blue line), between the predictions selected with DISOselect and the predictions generated by the most accurate disorder predictor at the dataset level, SPOT-Disorder (red line), and between the predictions selected with DISOselect and the best consensus-based method that relies on the support vector regression (SVR) (green line). Points indicate where the difference between protein AUCs crosses zero. The proteins are sorted by the value of the difference in the descending order. These results were published in [139]. 62

Figure 20: Chronological summary of the past surveys of the intrinsic disorder and intrinsic disorder function predictors. This figure was published in [165]. 66

Figure 21: Comparison of the predictive quality measured with AUC (panel A; solid lines) and MCC (panel B; dashed lines). We report results on the new benchmark (in green; dataset with <30% sequence similarity to the training proteins + with experimental validation of structured regions + with fully structured proteins), based on recent previous reports (in black; datasets with no limits on sequence similarity to the training proteins + with no experimental validation of structured regions + with only disordered proteins), and based on a similarity-limited benchmark (in red; a version of the new benchmark dataset with <30% sequence similarity to the training proteins + no experimental validation of structured regions + only disordered proteins). The latter dataset is a proxy for the datasets used in prior studies with the only difference being the reduced similarity to the training proteins. Disorder predictors are sorted by their AUC values on the new benchmark dataset. 72

Figure 22: Distribution of the AUCs (panel A) and MCCs (panel B) over nine disorder predictors. We exclude the poorly-performing GlobPlot from this analysis. The box plots show the lowest AUC (bottom error bar), first quartile (bottom of box), median (horizontal line inside box), third quartile (top of box) and highest AUC (top error bar). The grey plots are for the original datasets while the white plots are for the sampled/disorder content-equalized datasets that have similar distribution of the per-protein disorder content. The content distribution similarity was measured using Kolmogorov–Smirnov test at p-value of 0.001. 77

Figure 23: Comparison of the predictive quality measured with AUC (panel A; solid lines) and MCC (panel B; dashed lines). We report results on the generic set of disordered proteins (i.e., proteins that have disordered residues) from benchmark dataset (in black), the disordered protein-binding proteins (in yellow), and the disordered nucleic acids-binding proteins (in blue). Disorder predictors are sorted by their AUC values on the disordered proteins. 78

Figure 24: Prediction workflow of DisoLipPred. 86

Figure 25: Architecture of the deep recurrent neural network used by DisoLipPred. Panel A shows the partner-agnostic network that we train using the dataset of IDRs that interact with different partner types. Panel B gives the network that extends the partner-agnostic network to perform the partner-specific prediction of DLBRs. 89

Figure 26: Summary of the DisoLipPred’s predictions on the *Saccharomyces cerevisiae* proteome. Panel A shows the fraction of the yeast proteins predicted to have DLBRs. Panel B is the histogram of the putative content of DLBRs for the 4.9% of the yeast proteins with DLBRs.95

Figure 27: Analysis of the DisoLipPred predictions (Panel A) and the EspritZ-DisProt predictions (panel B) for the yeast proteins. The black arrows identify the rate of the putative proteins with DLBRs in the GO lipid associated protein set (i.e., set of 309 yeast proteins that share “lipid” keyword in the molecular function GO term and the “membrane” keyword in the cellular component GO term). Red lines show the distributions of the expected rates of the putative proteins with DLBRs, which we establish based on measuring the rate for 100 randomly selected sets of 309 yeast proteins. 96

Figure 28: DisoLipPred predictions for TatA protein (Uniprot: P69428; DisProt: DP00834). The blue line in the top panel shows the residue level propensity scores generated by DisoLipPred. The horizontal blue bars at the bottom are the corresponding experimental annotation of lipid binding regions and the binary prediction from DisoLipPred. The horizontal red bar shows the experimental annotation of the intrinsic disorder, where grey color identifies regions that lack disorder/order annotations. 99

List of Acronyms

AA: Amino acid

ASA:-Average accessible surface area

BLAST: Basic local alignment search tool

BUSCO: Benchmarking universal single-copy Orthologs

CASP: Critical Assessment of protein structure prediction

DLBR: Disordered lipid binding regions

DNA: Deoxyribonucleic acid

ELM: Eukaryotic linear motif

FPR: False-positive rate

GO: Gene ontology

IDP: Intrinsically disordered protein

IDR: Intrinsically disordered region

LR: Logistics regression

MCC: Matthews's correlation coefficient

MoRF: Molecular recognition feature

MSE: Mean Squared Error

NMR: Nuclear Magnetic Resonance

PCC: Pearson correlation coefficient

PDB: Protein data bank

PPR: Predicted positive rate

PR-AUC: Area under the precision recall curve

RNA: Ribonucleic acid

ROC-AUC: Area under the receiver-operating curve

SLiM: Short linear sequence motif

SVR: Support vector regression

TPR: True-positive rate

Abstract

COMPUTATIONAL ANALYSIS AND PREDICTION OF INTRINSIC DISORDER AND INTRINSIC DISORDER FUNCTIONS IN PROTEINS

By Akila Imesha Katuwawala

A dissertation submitted in partial fulfillment of the requirements for the degree of Engineering, Doctor of Philosophy with a concentration in Computer Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2021

Director: Lukasz Kurgan,
Professor, Department of Computer Science

Proteins, as a fundamental class of biomolecules, have been studied from various perspectives over the past two centuries. The traditional notion is that proteins require fixed and stable three-dimensional structures to carry out biological functions. However, there is mounting evidence regarding a “special” class of proteins, named intrinsically disordered proteins, which do not have fixed three-dimensional structures though they perform a number of important biological functions. Computational approaches have been a vital component to study these intrinsically disordered proteins over the past few decades. Prediction of the intrinsic disorder and functions of intrinsic disorder from protein sequences is one such important computational approach that has recently gained attention, particularly in the advent of the development of modern machine learning techniques.

This dissertation runs along two basic themes, namely, prediction of the intrinsic disorder and prediction of the intrinsic disorder functions. The work related to the prediction of intrinsic disorder covers a novel approach to evaluate the predictive performance of the current

computational disorder predictors. This approach evaluates the intrinsic disorder predictors at the individual protein level compared to the traditional studies that evaluate them over large protein datasets. We address several interesting aspects concerning the differences in the protein-level vs. dataset-level predictive quality, complementarity and predictive performance of the current predictors. Based on the findings from this assessment we have conceptualized, developed, tested and deployed an innovative platform called DISOselect that recommends the most suitable computational disorder predictors for a given protein, with an underlying goal to maximize the predictive performance. DISOselect provides advice on whether a given disorder predictor would provide an accurate prediction for a given protein of user's interest, and recommends the most suitable disorder predictor together with an estimate of its expected predictive quality. The second theme, prediction of the intrinsic disorder functions, includes first-of-its-kind evaluation of the current computational disorder predictors on two functional subclasses of the intrinsically disordered proteins. This study introduces several novel evaluation strategies to assess predictive performance of disorder prediction methods and focuses on the evaluation for disorder functions associated with interactions with partner molecules. Results of this analysis motivated us to conceptualize, design, test and deploy a new and accurate machine learning-based predictor of the disordered lipid-binding residues, DisoLipPred. We empirically show that the strong predictive performance of DisoLipPred stems from several innovative design features and that its predictions complements results produced by current disorder predictors, disorder function predictors and predictors of transmembrane regions. We deploy DisoLipPred as a convenient webserver and discuss its predictions on the yeast proteome.

Chapter 1. Introduction

Proteins are the actual workhorse behind the cellular functions in all living cells and viruses. These functions are governed by a mechanism known as the central dogma of biology where the genetic information encoded in DNA flows to RNA and where the messenger RNAs (genes) encode protein sequences [1]. Furthermore, the central dogma of structural biology is that one protein sequence has one structure and that this structure determines function of this protein sequence. Correspondingly, the conventional belief among biologists used to be that fixed and stable three-dimensional structures of proteins are essential for their function. Contemporary studies have revealed the existence of a large number of proteins and protein regions which perform important biological functions without having fixed and stable three-dimensional structures [2]. These regions without stable three-dimensional structure are generally identified as intrinsically disordered regions (IDRs). Proteins that include of one or more IDRs are referred to as the intrinsically disordered proteins (IDPs). The IDRs are widespread among three kingdoms of life as well as in viruses. Topical bioinformatics studies have suggested that IDRs are present in about 30% of proteins in Eukaryote, around 20% of proteins in Bacteria and Archaea domains and over 20% in viral proteins [3, 4]. A wide spectrum of important biological functions is performed by IDRs/IDPs. Interactions with biomolecules such as proteins, nucleic acids, lipids, metals etc., acting as domain linkers and entropic chains are some of common examples of these functions [5-7].

In general, computational tools contribute to protein studies from two basic viewpoints. First, they provide support to analyze massive amounts of data generated by high throughput protein sequencing platforms with the goal to reveal interesting patterns and trends. The other key contribution is their use as predictive platforms to explore untouched territories in protein sciences. The rate at which new protein sequences are being discovered has accelerated in recent years with the aid of new and cheap high-throughput sequencing techniques. Yet, compared to that rapid growth in the sequence space, we have access to only a relatively small amount of the experimentally annotated IDPs/IDRs, all the while studies suggest that they are prevalent in living organisms and viruses. The experimental annotations of IDPs/IDRs are deposited in public

repositories like Disprot[8] and Protein Data Bank (PDB)[9]. Conventional experimental techniques to annotate IDPs/IDRs are unable to keep up on par with the rate of production of the new protein sequences. As a result of this mismatch of rates, the gap between the number of the protein sequences and their available IDP/IDR annotations keeps widening. This is where the computational IDP/IDR predictors come into action. At this point, around 70 IDP/IDR related computational predictors were already published in the peer-reviewed literature [10-14]. Multiple studies show that they can accurately predict IDP/IDRs [15-18]. This work focuses on the computational disorder and disorder function predictions and assessment of these predictions.

1.1 Motivation

We investigate computational approaches that predict IDPs/IDRs and their related functions. We cover two basic themes, namely, predicting disorder and predicting functions of IDPs/IDRs in protein sequences. From the prediction of disorder perspective, to date the assessments of the predictive performance of disorder predictors have been carried out using large datasets without considering the performance at the individual protein level [14, 16]. We consider an interesting and new angle to investigate whether the datasets level performance is comparable to the performance at the protein level. Moreover, we note the number of the available computational methods that predict disorder in proteins. This large number of methods could confuse users, particularly those unfamiliar with this field. A proper guidance how to select suitable predictor would be invaluable to ensure that users can achieve highest predictive quality for a given protein of their interest. Considering the second theme related to predicting functions of IDPs/IDRs, to date there have been no study that investigates how well the current computational disorder predictors perform for specific functional subclasses of disordered proteins, particularly the commonly occurring IDRs that interact with proteins and nucleic acids. Such study would provide a valuable perspective for the development of predictors of specific functions of disorder. Moreover, some of functions in IDPs/IDRs do not have any methods that could be used to predict them in protein sequences [4, 18, 19]. The lipid binding is one such function of disorder that is

yet to be addressed with a predictive method. Correspondingly, we focus on addressing the following four issues:

1. How the predictive performance of disorder predictors at protein level compares with the current results that rely on the dataset level performance?
2. Is it possible to recommend predictors that perform well for specific protein sequences?
3. What is the predictive performance of the current disorder predictors for disordered proteins with protein binding regions and disordered nucleic acid binding regions?
4. Is it possible to accurately predict disordered lipid-binding residues from protein sequences?

1.2 Objectives and Proposed Contributions

This dissertation includes four main objectives which are organized in a sequential manner, where one objective flows into and motivates the next one. With the overarching goal to explore novel approaches for disorder and disorder function prediction we start with an exploratory analysis to identify relatively unexplored niche areas related to these predictions. This analysis motivates and informs us in the development of novel computational approaches to address specific issues in these niche areas.

The four main objectives are as follows.

Objective 1: Elucidation and comparative analysis of protein-level predictive performance for current disorder predictors.

We empirically investigate the disparity between protein-level and dataset-level predictive performance for the widely used disorder predictors. Since proteins are very diverse in their sequences and functions, we expect a considerable variation of the predictive performance across individual proteins, which is contrast to more “stable” and predictable benchmark dataset level performance.

Objective 2: Development of a novel protein-level predictor recommendation system to improve predictive performance of disorder predictions.

Findings from objective 1 show that certain methods perform particularly well (or rather poorly) for certain proteins. This predictive performance depends on the physiochemical properties of the individual proteins. We use this observation to conceptualize, design, build and test a novel recommendation engine that predicts the better performing methods for given individual proteins using their unique sequence-derived physiochemical properties.

Objective 3: Assessment and comparative analysis of the predictive performance of disorder predictions for specific functional types of disordered proteins.

We identify two common functional types of disordered proteins based on their binding partners: those that bind to proteins and nucleic acids. We assessed the predictive performance of widely used disorder predictors for above two functional subclasses of disordered proteins to investigate potential strengths and weaknesses of the current methods. This analysis aims to identify the functional subclasses of disordered proteins that may need further improvements in the quality of the disorder predictions. Motivated by the results from the objective 1, we perform this analysis at both dataset and protein levels.

Objective 4: Accurate prediction of the disordered lipid-binding residues from protein sequences.

Objective 3 reveals that predictions of disordered proteins that interact with proteins and nucleic acids are reasonably accurate. However, we note lack of tools that can predict interactions with other partner molecules, such as lipids. Consequently, we conceptualize, design, develop and comprehensively test a new computational tool that provides accurate prediction of the currently unexplored disordered lipid binding regions.

1.3 Organization of the dissertation

This section provides a brief outline to the flow of the upcoming chapters starting from chapter two.

Chapter Two provides background concerning the concepts and methods that are used in this research. As this is a multidisciplinary research area, the basic concepts cover two different

domains: Biological Sciences and Computer Science. Section one of the chapter two introduces the background on proteins and disorder in proteins. Furthermore, it describes the functional importance of intrinsically disordered proteins from a biological and biochemical perspective. The second part of this chapter introduces existing computational approaches to predict disorder and disorder functions in proteins.

Chapter Three and the following chapters are organized in a sequential manner to reflect the flow of the four objectives. In chapter three, we analyze the predictive performance of a comprehensive set of disorder predictors at the individual protein level. In this chapter, we contrast the protein level predictive performance of disorder predictors with their dataset level performance. Finally, we demonstrate the differences between dataset level and protein level performances through a detailed analysis and a case study.

Chapter Four is motivated by the findings from chapter three. Based on the differences in the predictive performances of disorder predictors at the individual protein level, we conceptualize, design and empirically test a novel system that recommends the most suitable/accurate disorder predictor(s) for individual proteins. This method aims to outperform the predictive quality offered by current representative set of disorder predictors. Moreover, it provides an innovative ability to estimate predictive performance of a given disorder predictors for a specific protein before the prediction is calculated.

Chapter Five analyzes a set of representative disorder predictors with respect to their predictive performance on two large functional subclasses of disordered proteins. We categorize disordered proteins based on their associated functions defined in the DisProt database. Next, we use the functionally annotated proteins from this database to perform comparative empirical assessment of the disorder predictors. We perform this assessment on multiple versions of the benchmark dataset after reducing the sequence similarity to the training datasets of evaluating predictors as well as after experimentally validating the unannotated regions in the disordered proteins. The motivation behind this analysis is to identify the areas that may require further improvements when it comes to the accuracy of the disorder predictions.

Chapter Six introduces a new methodology that predicts disordered lipid-binding residues. We review the current relevant methods to identify the fact that this important function of disorder lacks computational predictors. The current tools predict several other disorder functions and related transmembrane regions. We conceptualize, design and test an innovative deep recurrent neural network model that accurately identifies disordered lipid binding regions in proteins. We comprehensively validate predictive performance of this model and compare it to current related tools, with the goal to demonstrate that it provides high-quality and complementary results.

Chapter 2. Background and Related Work

2.1 Introduction to proteins, intrinsic disorder and its functions

Proteins are found in all living systems including simple organisms, such as bacteria/virus, as well as complex mammals, like humans. The word protein has originated from a Greek source *proteos*, which means “of prime importance” or “of first rank” [20]. As the name suggests, these are very (most) important biomolecules. The functions that proteins perform span across a wide range of biochemical and biological activities. Some of the examples are enzymes that catalyze biochemical reactions, hemoglobin for transportation, myoglobin for storage, and collagen for structure. The basic building block of proteins is called amino acid (AA). Several AAs bind to each other by peptide bonds to make polypeptide chains. One or more of these polypeptide chains folds in three-dimensional space to make a protein. The three-dimensional folding of polypeptide chains plays a crucial role to define the function for many of the proteins. The spatial conformation, commonly known as structure of the protein, is hierarchically organized into four levels. The primary structure is the linear sequential chains of amino acids, also referred as polypeptide chains. The secondary structure is the organization of the polypeptide chains into local and regular structural blocks that include α -helices and β -strands. The tertiary structure is defined by the 3D folding of those secondary structures. The highest level, the quaternary structure, involves spatial arrangement of multiple folded polypeptide chains.

2.1.1 Intrinsic disorder in proteins

The long-held convention is that proteins function comes as a linear flow from sequence to structure to the function. This notion emphasizes the fact that amino acid sequence governs the protein three-dimensional structure that directly determines the functions [2]. The three-dimensional structures of proteins are generally revealed through experimental techniques, like X-Ray crystallography, Nuclear Magnetic Resonance (NMR) and electron microscopy. The recently growing evidence reveals presence of a special class of proteins that have fundamentally challenged the classical paradigm that rigid three-dimensional structure is required for proteins

to function. These proteins exist as dynamic conformational ensembles without a fixed/rigid three-dimensional structure. They may lack the defined structure in entirety or in a specific region and correspondingly they are known as intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) [21-23]. The functional studies of IDPs/IDRs are done directly from the AA sequence, typically without attempting to reveal their (ensembles of) structure. IDPs/IDRs are estimated to be abundant in all living systems and viruses. Recent studies suggest that eukaryotes, bacteria and archaea have on average 19%, 6% and 4% of the disordered amino acids, respectively[4]. Furthermore, eukaryotic proteomes consist of 30% to 50% of proteins that have long IDRs (≥ 30 consecutive AAs) [3, 24]. The fully disordered proteins account for 6% to 17% of the proteins across various genomes[25]. The experimental annotations of IDPs/IDRs are deposited in several databases. Disprot [8] and IDEAL[26] include experimentally annotated information regarding function of IDPs/IDRs. Furthermore, while the Protein Data Bank (PDB) is primarily used as a repository of protein structures [9], is also provides information about IDPs/IDRs that “hide” in the regions with missing coordinates in the crystal structures and can be directly observed as extremely flexible residues in the NMR structures [27]. Even though these databases provide valuable information regarding IDPs/IDRs, they represent only a rather small number of proteins. The latest version of Disprot has 1 600 proteins and 3 500 IDRs, compared to the number of currently sequences proteins that has reached 214 million (as of April 2021).

Figure 1 illustrates the intrinsic disorder using the NMR structure of human Bcl-2-like protein 1 (PDB ID: 2ME8)[28]. This protein is an inhibitor for the programmed cell death (DisProt ID: DP00298)[29]. The three-dimensional structure of this protein was resolved by X-ray crystallography and NMR techniques. An intrinsically disordered region from 28th to 80th residue was identified by missing electron density of X-ray crystallography data[30]. Furthermore, experimental evidence shows that this intrinsically disordered regions acts as a flexible linker[30].

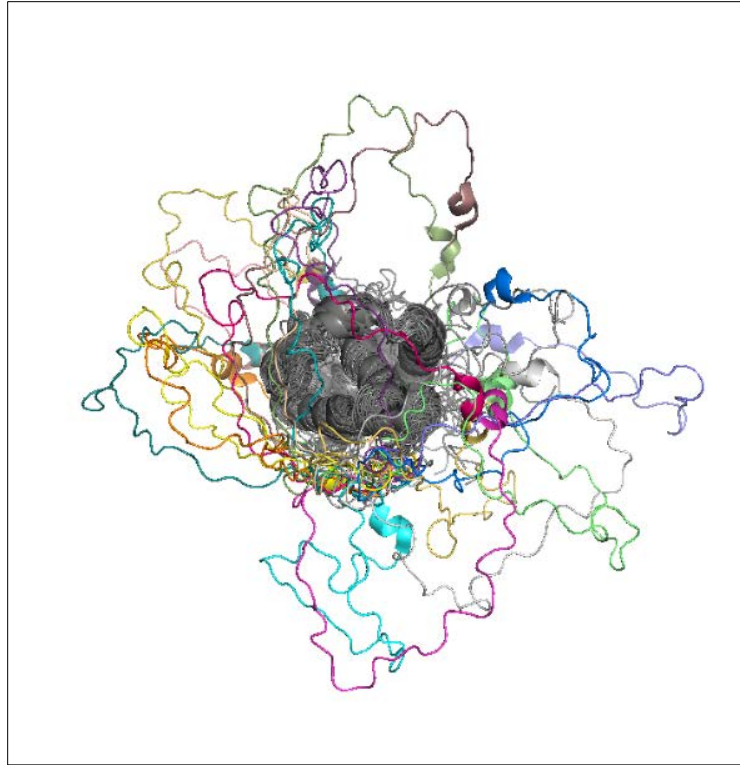


Figure 1: The PyMOL visualization of NMR structure of human Bcl-2-like protein 1 (PDB ID: 2ME8). All 20 different models are overlapped, and single color denotes a single model.

The PyMOL[31] cartoon visualization of the above proteins shows 20 different superimposed structures/conformation to illustrate the structural ensembles that make up disordered regions; each color denotes a single structure. The dense grey part in the middle corresponds to an ordered/structured region where all conformations converge into one fixed structure. The surrounding sparsely distributed multi-color sections corresponding to the disordered region.

2.1.2 Functional aspect of intrinsically disordered proteins

The wide range of functions that performs by IDPs/IDRs is broadly classified into two categories: by molecular level functions and by the interacting molecular partners [5, 22, 32]. This classification schema is followed by the DisProt database, the primary and largest source of the functional annotation of IDP/IDR [33, 34]. The molecular functions of the IDRs are further divided into seven classes: entropic chains, display sites, chaperons, effectors, assemblers, scavengers, and biological condensation[22]. Entropic chains are permanently unstructured to perform the

functions that needs a great degree of flexibility. IDRs present in the titin protein provides a good example for entropic chains[35]. Display sites act as facilitators for the post-translational modifications and they are located inside IDRs[36]. This localization facilitates downstream reactions that occur after catalytic site modifying enzyme approaches the effector protein[37]. The functional category of IDRs that acts as chaperons facilitates the folding of RNA molecules and other proteins into their functional conformations[38]. More than 33% of RNA chaperons and over 50% of protein chaperons are reported to be disordered[39]. The effectors show a functional shift upon binding to another molecule and transform from disorder to order state – this process is known as coupled folding and binding[40, 41]. The cell cycle regulatory proteins named p21 and p27 in the family of the cyclin dependent kinase and p53, which has multiple binding partners, are excellent examples of well-studied effectors. Assemblers serve as hub proteins that bring large number of proteins together to form large complexes. They act as either scaffolds or structural mortars to stabilize large molecular complexes [42, 43]. Examples of the assemblers that perform stabilizing function are the hub proteins which collects β -catenin, casein kinase I α , and glycogen synthetase kinase 3 β [44]. The scavengers digest and discard the debris of small ligands. Adrenaline and ATP scavenger by Chromogranin A are good examples for scavenging activity[45].

The other functional classification scheme of IDRs is according to their binding partners. This classification includes seven well-known categories of molecules that binds to IDR, namely proteins, DNAs, RNAs, lipids, metals, inorganic salt and small molecules. This supplements the information associated with some of the molecular function categories, such as effectors, chaperons, assemblers and scavengers.

The current version of 8.0 of the DisProt provides information for 2494 experimentally annotated unambiguous disorder regions. Table 1 shows the seven molecular level function level categories and their sub categories with the corresponding counts of proteins and functionally annotated disordered regions. The category with highest number of annotations is the molecular recognition assemblers, with 570 regions in 252 proteins. The other two highly annotated categories are the entropic chains (particularly the sub category of flexible linkers that has 175 regions) and molecular recognition effectors, which are covered by 448 and 488 regions,

respectively. Apart from above, the other four categories have relatively low numbers of annotations: recognition display sites with 270 regions, chaperones with 74 regions, scavengers with 54 regions, and biological condensation with 58 disordered regions in 40 proteins.

Table 1: Number of the molecular function annotations for the functionally annotated IDRs in the DisProt database. Annotations tagged as ambiguous have been excluded.

Functional annotations		Number of annotated IDRs	Number of annotated proteins
Molecular functions	Molecular function subcategories		
Molecular Recognition: Assembler	Assembler	100	51
	Localization (targeting)	20	11
	Localization (tethering)	38	22
	Total	570	252
Entropic Chain	Flexible linker/spacer	175	117
	Entropic bristle	12	6
	Entropic spring	3	2
	Self-transport through channel	4	2
	Structural mortar	1	1
	Total	448	283
Molecular Recognition: Effector	Inhibitor	94	59
	Activator	52	22
	DNA bending	4	3
	Disassembler	6	1
	DNA unwinding	1	1
	Total	488	237
Molecular Recognition: Chaperone	Space filling	8	2
	Entropic exclusion	6	3
	Total	74	33
Molecular Recognition: Display Site	Phosphorylation	76	53
	Glycosylation	10	5
	Fatty acylation	6	3
	Acetylation	14	10
	Ubiquitination	2	1
	Limited proteolysis	8	4
	ADP-ribosylation	1	1
	Methylation	9	7
	Total	270	169
Molecular Recognition: Scavenger	Metal binding/metal sponge	2	1
	Neutralization of toxic molecules	4	3
	Water storage	2	1
	Total	54	38
Biological Condensation	Liquid-liquid phase separation	12	12
	Amyloid	1	1
	Prion	16	7
	Total	58	40

Table 2 provides the annotation counts in terms of the binding partner-defined functions. The highest number of binding partner annotations are for the protein binding, with 1126 IDRs in 598 proteins. The binding partner annotations for other partners are lower and account for 143 regions for DNA binding, 71 regions for metal binding, 63 regions for RNA binding, 45 regions for lipid binding, 43 regions for small molecules, and just 3 regions for inorganic salt binding.

Table 2: Number of molecular partner annotations for the functionally annotated IDRs in the DisProt database.

Molecular Partner	Number of annotated IDRs	Number of annotated proteins
Protein	1126	598
DNA	143	88
Metal	71	43
RNA	63	50
Small molecule	43	35
Lipid	45	30
Inorganic Salt	3	3

2.2 Residue level protein structure and function prediction in disorder proteins

2.2.1 Computational prediction of intrinsic disorder in proteins

The interest in the annotation of intrinsic disorder proteins and ever growing gap between the available experimental annotations for disorder and the exponential growth in protein sequences motivates the development computational methods that predict disorder in protein sequences.

Figure 2 illustrates disorder and disorder function predictions. It shows results produced by three popular disorder predictors, IUPred-Long, IUPred-Short and SPOT disorder, for the human Bcl-2-like protein 1, which we introduced in section 2.1.1. Moreover, the disorder function predictor DFLpred was deployed to predict disordered flexible linker regions. Using the PyMOL's visualization of three-dimensional structure, the red color denotes ordered regions and blue color corresponds to a disordered region. The visualization of the predictions are generated using our recently released DEPICTER (Disorder Prediction Center) server [46], a new platform which provides simultaneous predictions of disorder and disorder functions. The results for each

predictor include numerical propensity scores, which quantify likelihood that a given amino acids is disordered (the top plot), and binary predictions, which categorizes each amino acid as either disordered or structured (presence of the horizontal bar corresponds to the disorder prediction).

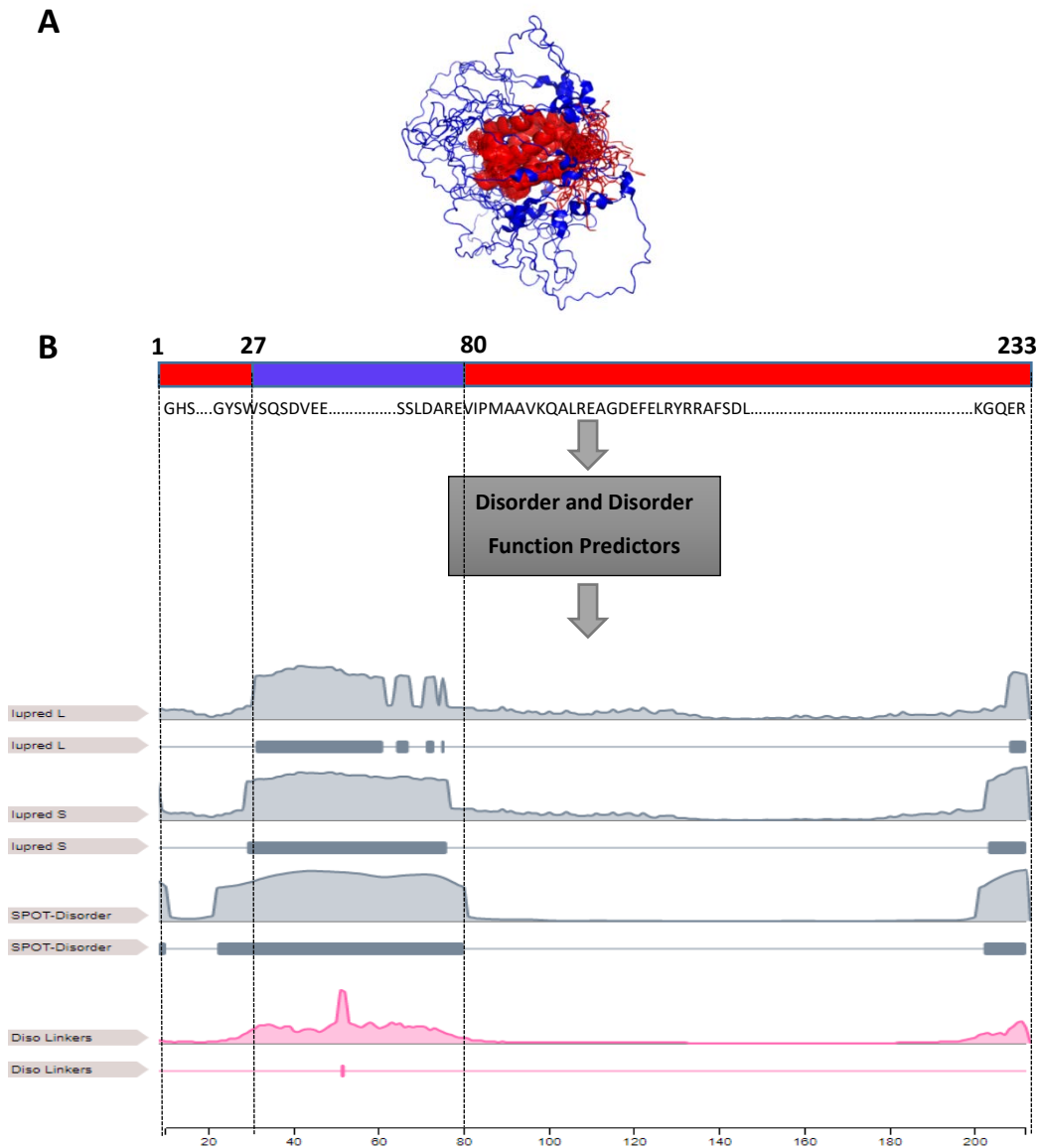


Figure 2: Panel A is three-dimensional structure of the human Bcl-2-like protein 1. The red and blue regions denote native (experimentally annotated) ordered and disordered regions, respectively. **Panel B** visualizes disorder and disorder function prediction from IUPred-Long, IUPred-Short, SPOT Disorder and DFLpred. The top horizontal bar gives the actual/native disorder and order annotations while the panels below show the putative propensities and the corresponding binary predictions (gray and pink horizontal bars) generated by the four predictors. The presence of a bar denotes presence of putative disorder or specific disorder function in the corresponding region.

The computational prediction of disorder benefits from the fact that amino acid sequence composition differs between the disordered and ordered regions [32, 47, 48]. Over 60 predictors of intrinsic disorder have been developed in the past forty years [11, 13, 16, 49]. They can be categorized in to three classes based on their underlying predictive models [11]:

1. Ab initio models.

These predictors rely on physical/analytical models that use physiochemical properties of amino acid sequences to differentiate between disordered regions and ordered regions. Some of the example predictors form this category would be NORSP [50] , GlobPlot[51] and IUPred [52].

2. Machine learning algorithms.

This category of predictors use machine-learning algorithms to produce predictive models. These models are trained on training set to maximize fit into the known disorder annotations and tested on an independent (low similarity to training set) datasets of proteins. This category includes a large number of predictors with some examples being RONN [53], DisEMBL [54], DISpro [55], DISOPRED [56], VSL2B [57], SPINE-D [58], SPOT-Disorder [59] .

3. Meta predictors.

Meta predictors generate predictions by combining outputs from several predictors with the intention of maximizing the predictive performance. The idea behind developing the meta predictors is to combine complementary disorder predictions to maximize their strong aspects (where they agree) and to minimize their weaknesses (where they disagree) [60]. Examples meta-predictors are MetaDisorder [61], MFDp [62, 63], disCoP [60], CSpritz [64], and ESpritz [65].

Furthermore, pre-calculated disorder predictions for about a dozen of popular methods are available in the MobiDB [66] and D²P² [67] databases.

The predictive performance of disorder predictors have been evaluated using several benchmark datasets. Many comparative surveys have been published [68-78]. These surveys evaluate predictive performance of the disorder predictors on different benchmark datasets that vary in

size from about 100 to over 20 000 proteins. One such popular assessment series was run as part of the Critical Assessment of protein Structure Prediction (CASP) experiments between 2002 (CASP5) and 2012 (CASP10) [68, 75-79]. A popular metric to evaluate predictive performance is the area under the curve of receiver operating curve (ROC-AUC). The ROC is a relation between true-positive rates (TPRs) and false-positive rates (FPRs) that is computed by using many thresholds to binarize numeric propensity scores produced by the predictors [80]. The AUC values range from 1.0 for perfect prediction and 0.5 for random prediction. The most accurate predictors have reached ROC-AUC of 0.89 [81] and 0.91 [79] in recent evaluations. These numbers suggest that disorder predictors perform with acceptable levels of predictive performance. Most of these predictors are available as standalone versions or webservers that are free to use for the scientific community.

Table 3 summarizes nine selected, highly-cited disorder predictors. The model type column provides information about the type of the predictive model used. Most of the disorder-predictors use machine learning-based models while only a few rely on other model types, like biophysics-inspired scoring functions. The total number of citations column provides a measurement of popularity for a given predictor. The annual citations give a measure that is more adequate for side-by-side comparisons, which quantifies how frequently a predictor was cited per year. We note that many of these methods are cited dozens of times every year, with the most cited predictors being disEMBL, IUPred, and DISOPRED3. The website column provides the link to the web resource where a given predictor can be downloaded or is available for online use. We observe that all highly-cited tools are available online.

Table 3: Summary of the nine highly-cited computational disorder predictors. The number of citations was collected from the Web of Science as of May 2020. The methods are sorted in the ascending chronological order.

Predictor Name	Year Published	Ref.	Model Type	Number of Citations	Annual number of citations	Website
ESpritz	2012	[65]	Bi-directional recursive neural network	185	23.1	http://protein.bio.unipd.it/espritz/
disEMBL	2003	[54]	Ensemble of feed-forward neural networks	840	49.4	http://dis.embl.de/
GlobPlot	2003	[51]	Derivative based curve optimization	678	40.4	http://globplot.embl.de/
JRONN	2005	[53]	Radial basis functional neural network	472	31.5	http://www.strubi.ox.ac.uk/RONN
VSL2B	2006	[57]	Support vector machine	480	34.3	http://www.dabi.temple.edu/disprot/predictor.php
IUPred	2009	[52]	Scoring function based on energy minimization	335	30.5	https://iupred2a.elte.hu/
DISOPRED3	2015	[56]	Ensemble of neural network, support vector machine and nearest neighbor	234	46.8	http://bioinf.cs.ucl.ac.uk/psipred/
DeepCNF	2015	[82]	Deep convolutional neural network	26	5.2	https://ttic.uchicago.edu/~wangsheng/software.html
SPOT-DISORDER	2016	[59]	Deep bidirectional neural network	73	18.3	http://sparks-lab.org/server/SPOT-disorder/

2.2.2 Computational prediction of functions in intrinsically disordered proteins

In parallel to the substantial efforts towards the development of a large number of computational disorder predictors, computational prediction of functions of IDRs has also gained a considerable amount of attention. The development of the disorder function predictors relies on the functional annotations from the DisProt database [33]. These annotations are used to design, optimize, and empirically test the predictive models. Over 20 computational methods that predict various functional traits of IDR were developed, implemented and published during the past decade [11]. These predictors predict disorder functions directly from the amino acid sequence of input proteins. Like in the case of the disorder predictions, the underlying predictive models use various types of machine learning algorithms and scoring functions [11]. Vast majority of these methods use data driven machine-learning models that are trained and tested using the available experimentally annotated functional IDRs. These methods attempt to minimize the predictive error on designated training dataset during training phase. After training is complete, they are tested on test datasets that include proteins that are explicitly dissimilar to the training proteins. The test proteins usually share low sequence similarity (<30%) against training dataset – this is to ensure that the evaluation is robust (overfitting training dataset does not lead to good results) and that the predictions cannot be simply performed using alignment-based methods.

The 23 disorder function predictors that were published during past decade fall under two main categories: 1) methods that predict IDRs that interact with specific binding partners; and 2) methods that predict molecular function of IDRs. The currently available methods cover three types of the molecular partners: proteins, DNA and RNA. The methods that predict molecular functions focus so far only on the flexible linkers (a type of entropic chains) and multifunctional (moonlighting) IDRs. Table 3 summarizes the 23 existing disorder function predictors. It shows a substantial interest towards developing new function predictors during past few years, as total of 17 predictors were published in the last 5 years (between 2015 and 2020) compared to 6 that were published prior to that (between 2007 and 2014).

Many of these disorder function predictors are implemented as publicly accessible web interface and/or publicly downloadable standalone software with the source code typically available in

freely accessible repositories. Table 3 reveals that 17 of the 23 predictors are accessible through their websites. Among these methods, 14 are available as webservers where predictions are done on the server side and can be obtained online. Moreover, 11 have downloadable source code that can be installed and run locally. Furthermore, six methods are available as both webserver and source code. We note that it is clear that the availability of these predictors is connected with the number of citations that they have received. The median annual number of citations for the methods without a webserver is 2 compared to 11 for the methods with webservers. The methods that provide only source code are cited at the annual median rate of 7 and predictors that have both code and webserver have received 9 citations per year. This speaks to the practical value for the availability of the webservers.

Table 4: Classification, citations and availability of the current predictors of IDR functions. The methods are *categorized* based on their predictive target (molecular partner vs. molecular function) and sub-type of the target (protein, DNA and RNA for molecular partners vs. flexible linker and moonlighting region for molecular functions). Predictors are sorted within each sub-type by the year of publication. The citations and availability are based on information as of Feb 25, 2019. The citations were collected using Google Scholar, where the annual citations are computed as an average number of citations per year since publication. Methods without any availability are listed as “not available” and those for which the websites cannot be found are denoted as “no longer available”.

Predictive target			Year	Method	Ref.	Citations		URL	
						Total	Annual		
Partners	Proteins	MoRFs	2007	alpha-MoRFpred	[83, 84]	445	37	Not Available	
			2010	retro-MoRFs	[85]	27	3	Not Available	
			2012	MoRFpred	[86, 87]	194	28	http://biomine.cs.vcu.edu/servers/MoRFpred/	
			2013	MFSPSSMpred	[88]	32	5	No Longer Available	
			2015	fMoRFpred	[89]	36	12	http://biomine.cs.vcu.edu/servers/fMoRFpred/	
			2015	DISOPRED3	[90]	206	52	http://bioinf.cs.ucl.ac.uk/disopred	
			2015	MoRFCHiBi	[91]	35	12	https://gspomerlab.msl.ubc.ca/software/morf_chibi/downloads/	
			2016	MoRFCHiBiLight	[92]	22	7	https://gspomerlab.msl.ubc.ca/software/morf_chibi/downloads/	
			2016	MoRFCHiBiWeb	[92]	22	7	http://morf.chibi.ubc.ca:8080/mcw/index.xhtml	
			2016	Predict-MoRFs	[93]	6	2	https://github.com/roneshsharma/Predict-MoRFs	
			2017	Wang et al. 2017	[94]	2	2	Not Available	
			2018	MoRFpred-plus	[95]	8	7	https://github.com/roneshsharma/MoRFpred-plus/wiki/MoRFpred-plus	
			2018	OPAL	[96]	8	6	http://www.alok-ai-lab.com/tools/opal/	
			2018	OPAL+	[97]	0	0	http://www.alok-ai-lab.com/tools/opal_plus/	
			2018	Fang et al 2018	[98]	0	0	Not Available	
	2019	Sharma et al. 2019	[99]	0	0	https://github.com/roneshsharma/BMC_Models2018/wiki			
			SLiMs	2012	SLiMPred	[100]	54	8	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/slimpred.php
				2016	PSSMpred	[101]	0	0	No Longer Available
			ALL	2009	ANCHOR	[102, 103]	388	39	http://anchor.enzim.hu
				2015	disoRDPbind	[104, 105]	44	11	http://biomine.cs.vcu.edu/servers/DisoRDPbind/
			2018	ANCHOR2	[106]	17	10	https://iupred2a.elte.hu/	
	DNAs		2015	disoRDPbind	[104, 105]	44	11	http://biomine.cs.vcu.edu/servers/DisoRDPbind/	
	RNAs		2015	disoRDPbind	[104, 105]	44	11	http://biomine.cs.vcu.edu/servers/DisoRDPbind/	
Functions	Flexible linkers		2016	DFLpred	[107]	17	8	http://biomine.cs.vcu.edu/servers/DFLpred/	
	Moonlighting regions		2018	DMRpred	[108]	0	0	http://biomine.cs.vcu.edu/servers/DMRpred/	

A vast majority of the available disorder function predictors are designed to predict protein-binding regions in IDRs (Table 4). This can be explained by the availability of the (by far) highest number of functionally annotated IDR in this functional category in DisProt (Table 2). The protein binding prediction methods can be further divided into specific type of protein binding IDR that they intended to predict. Molecular recognition features (MoRFs) are one such category that is targeted by 16 predictors. MoRFs are short protein binding regions, which usually span between 5 and 25 adjacent residues, that undergo disorder to order transition upon binding to their protein partners and which are placed inside of longer IDRs [41, 109]. The popular MoRF predictors include alpha-MoRFpred [110, 111], which predicts MoRFs that fold into alpha helices, MoRFpred [87], MoRFChiBi [91], fMoRFpred [41] and DISOPRED3 [112]. The second category of protein binding IDRs are short linear sequence motifs (SLiMs). They are short conserved motifs usually 3 to 12 amino acids long that are involved in protein interactions [113]. The Eukaryotic Linear Motif (ELM) resource provides a list of all currently known SLiMs that were used to design and test these predictors. SLiMpred [100] and PSSMpred [114] are the two methods that predict SLiMs. The remaining methods, which include ANCHOR [115], DisoRDPbind [116] and ANCHOR2A [52], predict a generic set of protein-binding IDRs, which include the short MoRF and SLiMs as well as longer protein-binding IDRs.

Computational approaches that predict functions other than protein binding are low in numbers. DisoRDPbind is the only method that predicts DNA binding and RNA binding IDRs. Moreover, DFLpred predicts flexible linkers and DMRpred predicts disordered multifunctional regions (moonlighting regions).

The popularity of the disorder function predictors can be quantified through the number of citations (Table 4). As of Feb 2019, The 25 predictors were cited 1651 times, with the median number of citations at 22. Based on the annual citation numbers, the most popular predictors are DISOPRED3 (52 citations per year), ANCHOR (39 citations per year), alpha-MoRFpred (37 citations per year), MoRFpred (28 citations per year), and fMoRFpred and MoRFChiBi (12 citations per year). The high number of citations for DISOPRED3 could be partly due to its ability to predict disorder in general in addition to prediction of protein binding regions in IDR.

2.2.3 Computational modelling related to the development of disorder predictors

Prediction of IDRs and their functions in proteins can be viewed as a classification task from the machine learning perspective. The classification task requires a fixed-size input. In this case, the raw input is the amino acid sequence of proteins. Since these amino acid sequences are represented as a text string of variable length, they must be converted into a fixed-length numerical feature vector. This is where physiochemical and putative structural properties of amino acids are used to encode sequences into the feature vectors. For example, sequence can be converted into 20-dimensional vector of the frequencies of the 20 amino acids. Next, the input feature vectors are processed by predictive models to produce predictions of disorder and disorder functions. Various machine learning algorithms are used to produce predictive models. Popular examples include Logistic Regression [117], Decision Trees [118], Naïve Bayes [119], Random Forest [120], and (recently) Deep Neural Networks [121]. The training and optimization of the machine learning models (including in some case feature selection) is typically done through cross validation on the training datasets, with the intention to minimize overfitting. Once the models perform adequately well on the training datasets, their predictive performance is validated using independent (sharing low similarity with the training proteins) test datasets.

The dataset definition and preparation play a crucial role to instill trust for the built predictors. The usual practice is to divide the currently available data into two parts: training set and test set. Further, it is important to ensure that training dataset and test datasets are dissimilar from each other, as measured by the protein sequence similarity. This similarity is usually reduced to less than 30% between training set and test set. This way the test proteins cannot be accurately predicted from the training proteins using sequence alignment. The training dataset is further divided into several folds for the cross validation. In the cross validation, the training dataset is divided into equally sized (in terms of number of proteins) x subsets (folds), and in the i^{th} ($1 \leq i \leq x$) fold of the cross validation, $x-1$ subsets are used to train the model, and the remaining i^{th} subset is used as test set to evaluate the trained model. The results of the cross validation tests are reported as the aggregate over all test folds or as an average over the x folds of tests.

2.2.4 Evaluation of predictive performance

We use numerical evaluation criteria/metrics to assess the predictive quality of disorder and disorder function predictors. These predictors typically generate two forms of outputs. The first is the real-valued propensity for disorder (disorder function) that quantifies likelihood that a given amino acid is disordered (has disorder function). The second is the binary prediction where 0 usually means ordered (lacking given function) and 1 means disordered (having given function). In fact, in most cases the binary prediction is generated from the propensities, such that residues with propensities greater than or equal than a given threshold are predicted as disordered (functional); otherwise they are predicted as ordered (non-functional).

One of the most common metrics that is used to evaluate the propensities is the area under the receiver-operating curve (ROC-AUC). The ROC is a relation between true-positive rates (TPRs) and false-positive rates (FPRs) that is computed by using many thresholds. Typically, the thresholds are set to equal the set of all unique propensities produced by a given predictor. TPRs and FPRs are calculated by comparing the native annotation of disorder/disorder function with the predictions at different thresholds. TPR and FPR are defined as:

$$\text{TPR} = \frac{TP}{TP + FN} = \frac{TP}{\text{number of all native functional residues}}$$

$$\text{FPR} = \frac{FP}{FP + TN} = \frac{FP}{\text{number of all native non - functional residues}}$$

where TP is the number of true positives (correctly predicted positives), FN is the number of false negatives (positive residues that are predicted as negative), FP is the number of false positive (negative residues that are predicted as positive), and TN is the number of true negatives (correctly predicted negatives). Given TPR and FPR values generated at different thresholds ranging from 0 and 1, we plot the ROC curve and calculate the corresponding AUC value. Note that positive means disordered (having a given disorder function) while negative means ordered (not having a given disorder function).

We note that the classification in the context of disorder/disorder function prediction is imbalanced. Significant majority of residues are ordered/lack specific function. Thus, it is

desirable to evaluate in the regime where we set a low false positive rate (FPR), e.g. at or below 5% or 10%. This ensures that the predictors do not over-predict the disorder/disorder function. Correspondingly, we calculate AUC_{lowFPR} that covers the ROC curve for the low range of FPR values (typically between 0 and 5%). Since AUC_{lowFPR} are rather small and difficult to assess directly, we compute $AUC_{ratio} = AUC_{lowFPR}/AUC_{random_lowFPR}$, where AUC_{lowFPR} is divided by the AUC of a random predictor (for which FPR always equals to TPR) in the same FPR range. This ratio quantifies the rate of improvement over a random predictor, i.e., ratio > 1 means that a given method is better than random and ratio of two indicates that this method is twice better than random.

In addition to the calculating the AUC_{lowFPR} , another relatively popular measure to calculate the predictive performance for an imbalanced dataset is the area under the precision recall curve (PR-AUC). The PR-AUC is a functional relation between the precision and recall values computed by using thresholds that binarize the propensity scores generated by the predictors. Higher precision and higher recall values means lower false positive rate and lower false negative rate respectively. This means that higher PR-AUC values correspond to more accurate predictions.

To evaluate the binary predictions, we use accuracy, precision, sensitivity and Matthews Correlation Coefficient (MCC):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{\text{number of all residues}}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{\text{number of all predicted functional residues}}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{\text{number of all native functional residues}}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The AUC ranges between 0.5 (equivalent to a random prediction) and 1 (perfect prediction). Accuracy, precision and sensitivity range between 0 and 1 where 0 denotes that no residues were predicted correctly and 1 denotes perfect prediction. MCC ranges between -1 and 1, where -1

denotes that the prediction is inverted (all functional residues are predicted as non-functional residues and vice versa), 0 denotes a random result and 1 denotes the perfect prediction.

We also emphasize the importance to measure the consistency of differences in the predictive performance between two predictors, i.e., to assess statistical significance of the differences. We place the n results from the two predictive models (e.g. using n test proteins) side by side, and have n pairs of results. We use the Student's paired t -test if both sets of the n results follow normal distribution, and otherwise we use Wilcoxon signed-rank test. Student's paired t -test [122] evaluates two groups of data (e.g., two groups of AUC values), and determines whether their mean values are significantly different. Wilcoxon signed-rank test [123] is an alternative to student's paired t -test when X_1 or X_2 does not follow normal distribution. It determines the difference between the medians of X_1 and X_2 . We use Anderson-Darling test (at the 5% significance) to verify if a sample of n results follows the normal distribution. The Anderson-Darling test [124] is a statistical test that checks whether a set of data follows a certain probability distribution.

Chapter 3. Elucidation and comparative analysis of protein-level predictive performance for current disorder predictors

This chapter describes work related to objective 1 from section 1.3. The results and methods reported in this chapter were published in [125].

Large-scale evaluation of predictive quality for computational disorder predictors have received substantial amount of attention [68-78]. These evaluations usually report popular metrics such as area under receiver operating curve (ROC-AUC), accuracy, sensitivity, specificity, and MCC. The usual practice is to evaluate these predictors at a dataset level without considering the performance at the individual protein level. However, users typically apply these predictors to individual proteins rather than large datasets.

This chapter investigates the predictive quality of selected disorder predictors at individual protein level, besides the usually considered dataset level performance. We also contrast our results with a few previously done dataset level assessments to ensure that our results cross-check with these studies and to provide context for the protein level analysis. The first part describes details of the benchmark dataset that we used and selection criteria that we used to derive the list of the considered disorder predictors. Then we assess the selected set of computational disorder predictors at both dataset and protein levels. The next section investigates complementarity/similarity of predictive quality between the selected set of predictors. The complementarity study intends to demonstrate differences and similarities between predictive performance and the potential to use these methods in tandem. Finally, the chapter presents a case study using a selected protein to demonstrate the predictions coming from several different methods.

3.1 Related work in disorder predictor assessment

Several comparative assessments have been carried out in recent years to benchmark the predictive quality of computational disorder predictors [15, 17, 18, 68-78, 126]. These assessments compare the predictive performance of a given set of disorder predictors on a

benchmark dataset. These benchmark datasets vary in size between 100 and 20 000 proteins. The critical assessment of protein structure predictions (CASP) is one such bi annual assessment, which considered disorder predictions from 2002 (CASP5) to 2012, (CASP10) [68, 73, 75-78]. CASP experiments have used relatively small datasets with around 100 proteins. Some of the more recent assessments rely on datasets from CASP10 experiment[13], use a larger datasets of 250 proteins collected from DisProt [18] and 500 proteins collected from PDB and DisProt[15]. The evaluation with the largest dataset covers 13 predictors evaluated on a dataset of over 25 000 proteins[126]. Furthermore, a set 13 computational disorder predictors was assessed on dataset of 350 membrane proteins [127]. A comprehensive survey of these assessments was done recently in [128].

The results that are reported in the above assessments can be used to evaluate the relative performance of the computational disorder predictors at the dataset level. None of the currently available surveys focused on evaluating the disorder predictors at the individual protein level. This is a noteworthy concern as arguably most of the time users characterize disorder for individual protein rather than large-scale datasets. Many instances of the protein-level analysis can be found in literature. For example, MFDp predictor[129], which was developed in our lab, was used to predict disorder in flagella capping protein[130], Cia2[131], SpSM30B[132], AP24[133] and BRCA1[134], among many other proteins.

We use Figure 3 to illustrate differences between the protein and the dataset level analyses. the figure contrasts the dataset level predictive performance reported for few selected computational disorder predictors in the recent large-scale assessment [126] with predictive performance for a few selected proteins. The selected disorder predictors include three methods that are characterized by different ranges of the dataset level performance: GlobPlot on the lower end, IUPred for medium performance, and VSL2b for high end, according to the considered assessment. The evaluation measure is ROC-AUC that ranges from 0.5 to 1.0. The reported dataset level performance for GlobPlot, IUPred and VSL2b are 0.631, 0.726 and 0.821 respectively. The individual predictive performance for the proteins such as guanylate kinase (red marker in Figure 3; UniProt Q2GLF7) and phytoene desaturase (blue marker; UniProt P21685) are in sync with the datasets-level AUCs. At the same time, for nonhemagglutinin type D (green

marker in Figure 3; UniProt Q9LBR2) the protein level performance varies substantially from the dataset level performance. The GlobPlot’s performance for this protein is considerably better than its dataset level performance while the results of the other two predictors are similar to their dataset-level performance. In this case, the worst at the dataset level GlobPlot outperforms the other two methods. Another interesting example is hydroxymethyltransferase (orange marker in Figure 3; UniProt P0A5Q8) where the highest predictive performance is provided by IUPred, while the dataset-level assessment shows that VSL2B outperforms IUPred. Overall, all three predictors outperform their corresponding dataset-level results for this protein. The alkaline phosphatase protein (violet marker; UniProt A1YYW7) is predicted poorly by both GlobPlot and IUPred, at level much lower than their dataset-level benchmark suggests, while the predictions from VSL2B substantially outperforms its dataset-level AUC.

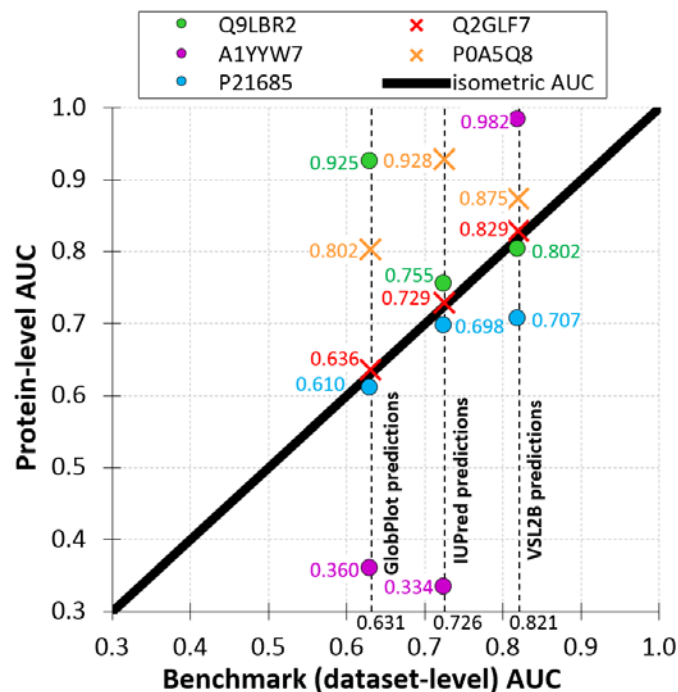


Figure 3: Comparison of the benchmark (dataset-level) and protein-level predictive performance measured with AUC for three disorder predictors (GlobPlot, IUPred and VSL2B) and five color-coded proteins: non-toxic nonhemagglutinin type D (green marker; UniProt Q9LBR2), guanylate kinase (red marker; UniProt Q2GLF7), alkaline phosphatase (violet marker; UniProt A1YYW7), hydroxymethyltransferase (orange marker; UniProt P0A5Q8) and phytoene desaturase (blue marker; UniProt P21685). The benchmark AUC values are shown on the x-axis, while the protein-level AUC values are color-coded, shown on the y-axis and their values are given next to the corresponding markers. The black isometric AUC line shows equivalent dataset-level and protein-level values. Published in [125]

Altogether, Figure 3 clearly illustrates that protein level predictive performance can vary considerably from the dataset level results that are reported by previous assessments. This chapter assesses and investigates variability of the protein level predictive performance, and contrasts these results with the dataset level performance for a set of 13 representative disorder predictors using a large dataset of 6,271 proteins.

3.2 Benchmark datasets and selection of disorder predictors

This evaluation was done on a protein dataset with disorder annotations that was originally used in the large prior dataset-level assessment [126]. The initial dataset in that study included 25,717 proteins for which annotations, disorder predictions and protein sequences were extracted from the MobiDB database [66]. We improved the original dataset to remove low quality protein sequences and to reduce redundancy, i.e., the datasets includes many clusters of similar proteins that could skew results toward these over-represented protein families. We removed sequences with unknown/undetermined amino acid (AA) types, which is required to generate disorder predictions, and by reducing within-dataset redundancy. We used BLASTCLUST[135] to cluster proteins with over 25% similarity and we selected one protein from each cluster. The final dataset has 6,271 proteins that share <25% similarity and that include 105,709 disordered and 1,672,907 structured residues. This dataset was also utilized in another study that investigates quality assessment of disorder predictions [136]. This study shows that the predictive performance of the disorder predictors on the original dataset of 25,717 proteins that was assessed in [126] is very similar to the performance on the improved benchmark dataset with 6,271 proteins [136].

We selected a diverse set of 13 publicly available computational disorder predictors for our analysis. Ten out of the 13 predictors were used in the previous large scale assessment [126]. The selected 13 predictors include three versions of ESpritz that predict intrinsic disorder annotated from x-ray structures (ESpritz-Xray), NMR structures (ESpritz-NMR) and using DisProt database (ESpritz-DisProt)[65]; the two versions of IUPred that are separately optimized to predict short IDRs (IUPred-short) and to predict long IDRs (IUPred-long)[52]; the two versions of DisEMBL which are developed for X-ray structures (DisEMBL-465) and to predict IDRs that form loop secondary structures (DisEMBL-HL)[54]; GlobPlot[51], RONN[53] and VSL2B[57]. We excluded

three methods SEG [137] , Pfilt [138] and FoldIndex [139] that were used in the previous assessment, considering that they are rather old and that their dataset level predictive performance is low [126]. Furthermore, we added three recent disorder predictors: DISOPRED3 [112] (that performed very well in CASP10 [73]) and two recently published deep learning-based methods: SPOT-Disorder [112] and DeepCNF-D [82]. We use the fast version of DeepCNF-D (DeepCNF-D ami_only) given the large size of our dataset. These methods were published between 2002 and 2016 (work presented in this chapter was done in 2017 and 2018) and most of them use machine learning-based predictive models. They are well-cited, with the annual number of citations ranging between 5 (for the new methods) and 50. The selected predictors uniformly cover the three categories of methods (Section 2.2.1) including *ab-initio* tools (IUPred-short, IUPred-long and GlobPlot), machine learning-based predictors (RONN, DisEMBL-HL, DisEMBL-465, VSL2B, DeepCNF-D and SPOT-Disorder) and meta-predictors (DISOPRED3, ESpritz-Xray, ESpritz-NMR and ESpritz-DisProt). They were designed to address prediction of all major types of disorder annotations including annotations that rely on x-ray crystal structures, NMR structures and a variety of other experimental methods that are covered in the DisProt resource.

Table 5: Dataset- and protein-level predictive quality for the 13 considered disorder predictors. The dataset-level accuracy and AUC are compared against the previously published results. We note that false positive rates are typically not reported in the past studies. Protein-level results are summarized with the median value. The methods are sorted by their dataset-level AUC on our benchmark dataset. These results were published in [125].

Disorder predictor	Accuracy (binary predictions)				AUC (putative propensities)				FPR (binary predictions)	
	Protein-level median	Dataset-level	Previously reported dataset-level	Difference dataset-level	Protein-level median	Dataset-level	Previously reported dataset-level	Difference dataset-level	Protein-level median	Dataset-level
GlobPlot	0.876	0.855	0.847	0.8%	0.662	0.626	0.631	0.5%	0.090	0.111
disEMBL-HL	0.715	0.713	0.721	0.8%	0.780	0.725	0.727	0.2%	0.282	0.277
IUPred-long	0.945	0.922	0.921	0.1%	0.798	0.732	0.726	0.6%	0.012	0.040
JRONN	0.848	0.847	0.839	0.8%	0.824	0.772	0.759	1.3%	0.132	0.131
ESpritz-NMR	0.931	0.905	0.903	0.2%	0.879	0.776	0.770	0.6%	0.041	0.068
IUPred-short	0.934	0.921	0.924	0.3%	0.852	0.778	0.778	0.0%	0.043	0.051
disEMBL-465	0.931	0.921	0.925	0.4%	0.835	0.780	0.787	0.7%	0.047	0.049
ESpritz-Xray	0.904	0.849	0.840	0.9%	0.904	0.796	0.778	1.8%	0.071	0.136
VSL2B	0.832	0.816	0.805	1.1%	0.874	0.810	0.821	1.1%	0.161	0.177
ESpritz-DisProt	0.959	0.917	0.934	1.7%	0.861	0.816	0.791	2.5%	0.000	0.034
DeepCNF	0.952	0.936	0.944	0.8%	0.926	0.871	0.898	2.7%	0.027	0.033
DISOPRED3	0.977	0.957	0.955	0.2%	0.969	0.899	0.897	0.2%	0.004	0.016
SPOT-Disorder	0.971	0.956	0.950	0.6%	0.969	0.904	0.891	1.3%	0.011	0.018

3.3 Predictive performance of disorder predictors

3.3.1 Dataset level predictive performance

Table 5 provides dataset level predictive performance of 13 disorder predictors on the benchmark dataset with 6,271 proteins. Furthermore, this analysis contrasts the dataset level performance with previous assessments. The previously published performance for the DISOPRED3 comes from CASP10 [79], for SPOT-Disorder from [59] and for DeepCNF from [82]. The other ten predictors were assessed in [126]. The “difference from dataset level” columns show the percentage of the differences between the previously reported results and the assessment on our dataset for the accuracy and AUC measures. The average differences (over the 13 predictors) in accuracy and AUC are 0.67% and 1.04%, respectively. This shows that our analysis that uses a larger dataset closely reflects the current state of the per-dataset assessments. The results are also in agreement with the previous assessments showing that DeepCNF, SPOT-Disorder and DISOPRED3 outperform the other 10 disorder predictors.

3.3.2 Protein level predictive performance

The analysis of the protein level predictive performance explores the per protein distributions of three commonly used criteria of the predictive performance: AUC, accuracy (ACC) and false positive rate (FPR). Figure 4 illustrates distribution of these metrics over the 6,271 proteins. The distributions of the accuracy and AUC metrics for all predictors are right-skewed with long tails. The corresponding distributions of the FPR values are left-skewed with similarly long tails; this is because larger values of FPR indicate lower predictive quality. These distributions demonstrate that while majority of the proteins are predicted with above average predictive performance, minority of proteins that are located in the long tails are predicted with low performance.

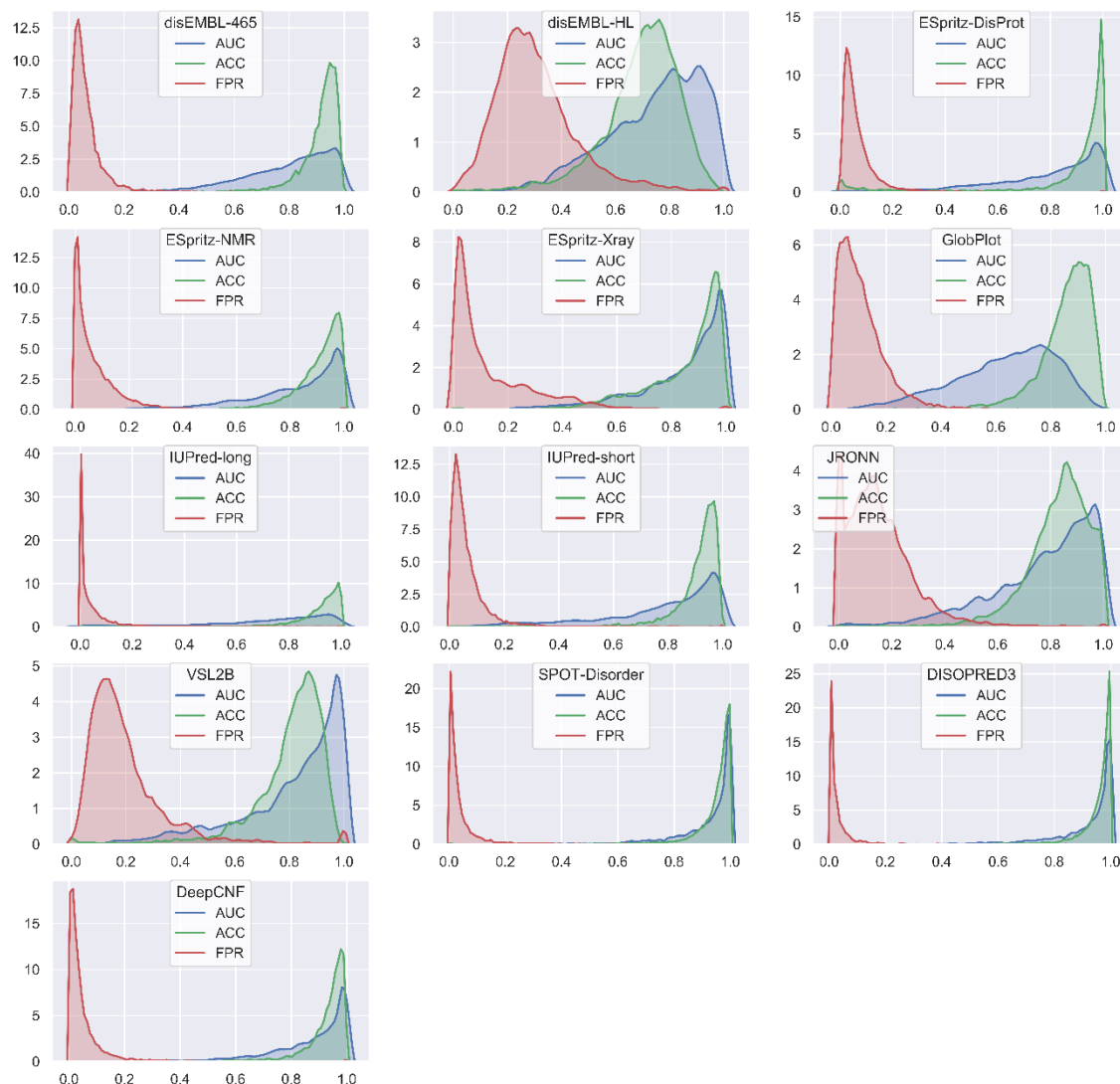


Figure 4: Distributions of the protein-level predictive quality measured with accuracy (green plots), AUC (blue plots) and false positive rate (FPR; red plots) for the 13 disorder predictors. The y-axis gives the fraction of the proteins in a given range of accuracy/AUC/FPR values. These results were published in [125].

The above distributions are summarized and compared with the corresponding dataset level results in Figure 5. Panels 5A, 5B and 5C show accuracy, ROC-AUC and FPR, respectively. The box plots show the first quartile (in red), second quartile (median; where red and green meet) and third quartile (in green), with whiskers that denote the 10th and 90th percentiles. The long tails are represented by the long bottom whiskers for accuracy and AUC (long top whiskers for FPR) when compared to the corresponding top whiskers (bottom whiskers for FPR). The black horizontal lines denote the dataset-level values. Figure 5 reveals that majority of the proteins

secure higher levels of predictive performance than their corresponding dataset-level assessment suggests. Furthermore, protein-level medians for accuracy and AUC are consistently higher or at worst similar to the dataset-level values, while the protein-level medians for FPR are consistently lower or similar when compared to the dataset-level values. These values can be directly compared in Table 5. This trend is consistent across the 13 disorder predictors and the three measures of predictive performance.

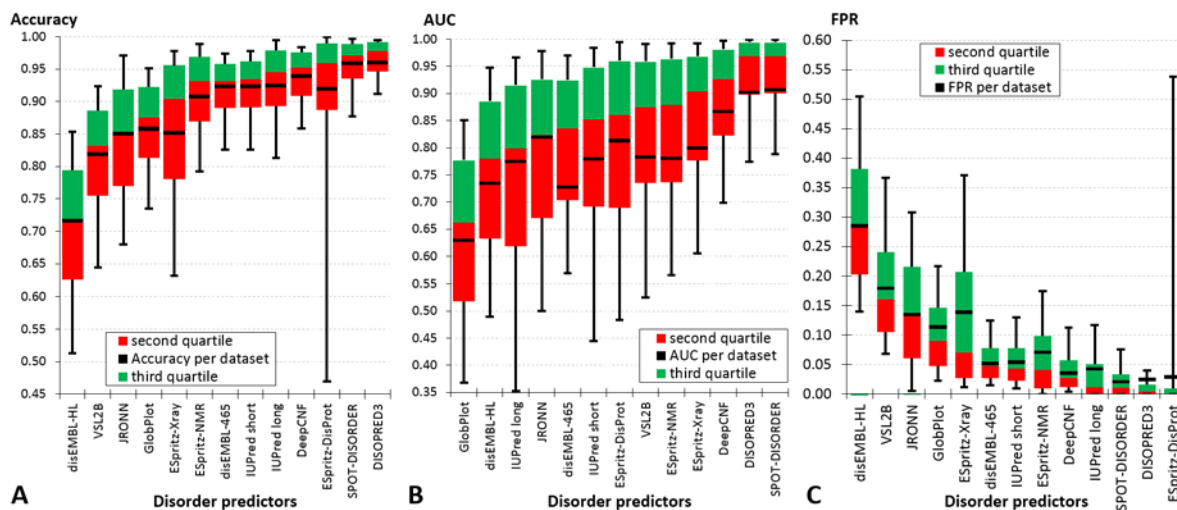


Figure 5: Distributions of the protein-level predictive quality measured with accuracy (panel A), AUC (panel B) and false positive rate (panel C) for the 13 disorder predictors. Box plots show the second quartile (in red), median (between red and green boxes) and third quartile (in green) for the distribution of the protein-level values. The whiskers denote the corresponding 10th and 90th percentiles. The black horizontal lines show the benchmark dataset-level performance. The predictors are sorted by their median values of the predictive performance. These results were published in [125].

We observe that the fraction of proteins for which performance is better than the dataset level estimate varies between the disorder predictors. Proteins for which a given predictor performs better than expected (i.e., better than its dataset level estimate) are denoted as easy-to-predict proteins and proteins which it performs lower than expected (i.e., lower than its dataset level estimate by a margin equal to the average difference of the pairwise comparison of all predictors) are denoted as hard-to-predict proteins. The value of the margin equals 0.067 and 0.071 for accuracy and AUC, respectively. Figure 6 illustrates the rates of the hard-to-predict and easy-to-predict proteins for each disorder predictor evaluated with AUC and accuracy (ACC).

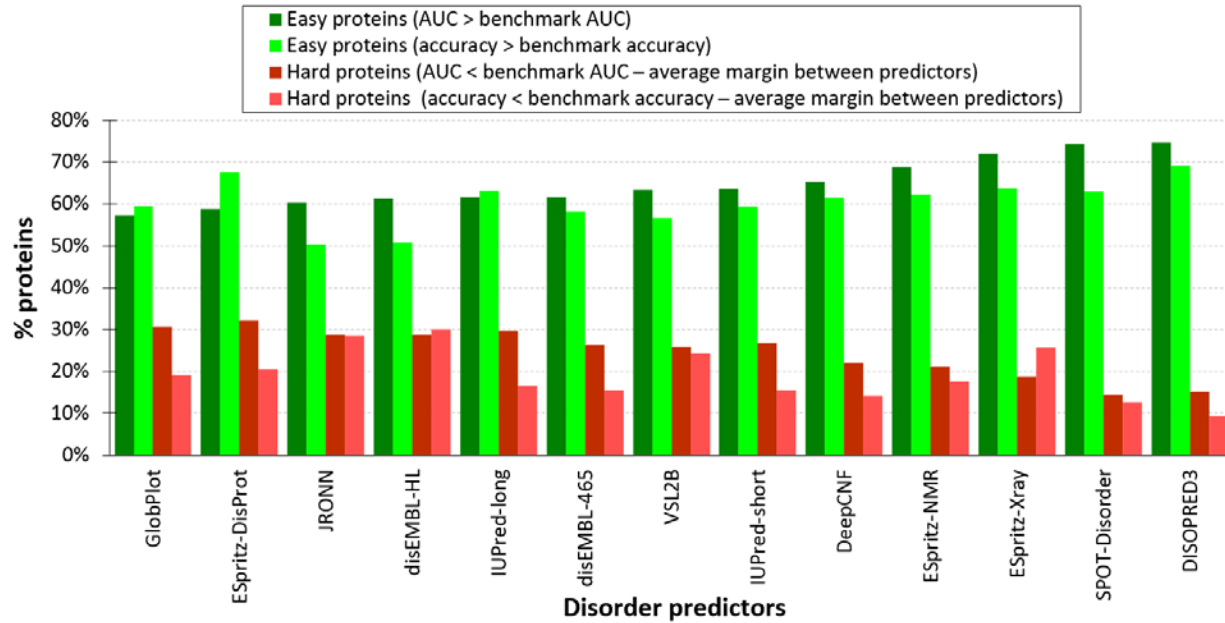


Figure 6: Analysis of the easy- and hard-to-predict proteins for the 13 disorder predictors. The easy proteins are predicted with higher-than-expected accuracy or AUC, i.e. their protein-level accuracy (AUC) > dataset-level accuracy (AUC). The hard proteins are predicted with relatively low accuracy or AUC, i.e. their protein-level accuracy (AUC) < (dataset-level accuracy (AUC) – average margin of difference between disorder predictors). Bars represent the fraction of the easy proteins (green bars) and the hard proteins (red bars) when predictive performance is quantified with AUC (dark shade) and accuracy (light shade). Predictors are sorted by fraction of the easy proteins quantified with AUC (dark green bars). These results were published in [125].

When it comes to AUC, between 57% of proteins (for GlobPlot) and 75% of proteins (for DISOPRED3) are easy to predict, i.e., their predictive performance higher than the expected value. Similarly, between 50% (for JRONN) and 69% (for DISOPRED3) of proteins have better than expected accuracy. Red bars in Figure 6 quantify the abundance of the hard-to-predict proteins. Figure 6 shows that between 14% of proteins (for SPOT-Disorder) and 32% of proteins (for ESpritz-DisProt) are hard-to-predict with respect to their AUCs. Similarly, the analysis reveals that between 9% (for DISOPRED3) and 30% (for disEMBL-HL) of proteins have lower-than-expected accuracy. This means that the users should expect low-quality protein-level predictions for anywhere between 10% (for accurate predictors like DISOPRED3 and SPOT-Disorder) and 30% (for less accurate predictors like JRONN and disEMBL-HL) of proteins that they submit. In other words, the end users will be unpleasantly surprised with their results for anywhere between 1 in 10 proteins to 3 in 10 proteins.

A few studies have reported that predictive performance of disorder predictors is sensitive to the native protein-level disorder content [17, 59, 126]. The trend that they observe is that performance on proteins with a substantial amount of disorder is lower compared to the proteins with little to no disorder. We investigate this trend by comparing the protein-level accuracy against the protein sequence length and fraction of the disordered residues. The Pearson correlation coefficients (PCCs) computed for accuracy vs chain length for the 13 disorder predictors are low and range between 0 and 0.37, with the average of 0.12. In contrast, the PCCs between disorder content and accuracy are relatively high and vary between -0.14 and -0.75 , with the average of -0.42 . The negative sign implies that, as expected, proteins with more disorder score (on average) lower predictive performance.

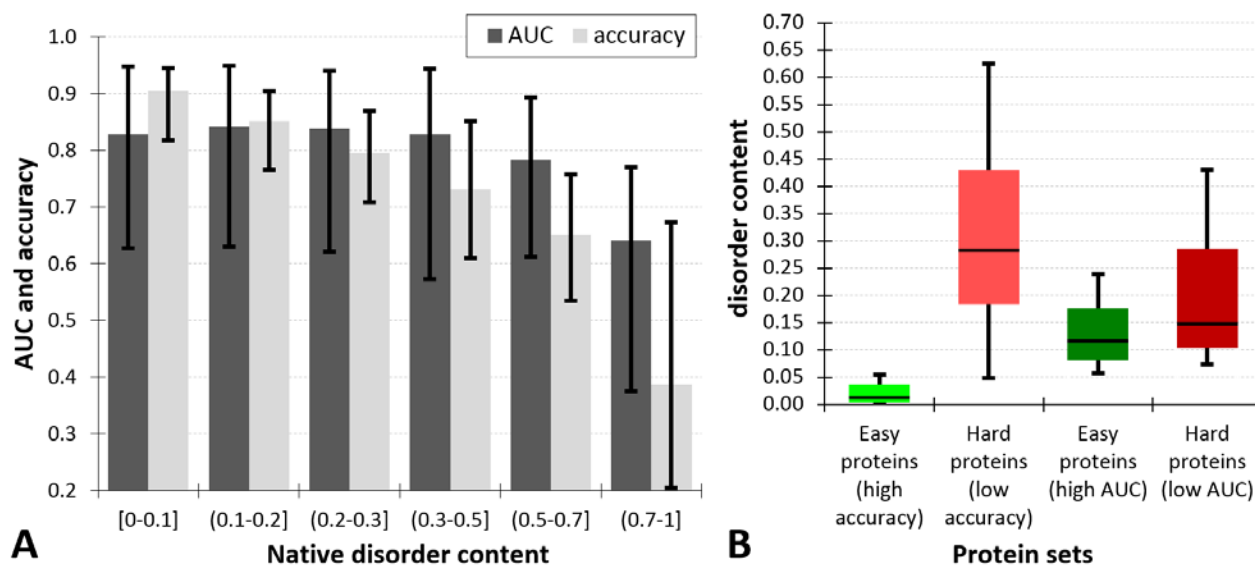


Figure 7: Relation between the protein-level predictive performance and the native disorder content. Panel A shows medians of the average (over the 13 predictors) accuracy and AUC for proteins grouped by their native disorder content, defined as the fraction of disordered residues in the sequence. The whiskers give the 10th and 90th percentiles of these averages. Panel B gives the distribution of the disorder content for the easy and hard proteins that are in common across the 13 predictors. The box plots show the 2nd quartile, median (black horizontal line) and 3rd quartile for the distribution of the protein-level disorder content values. The whiskers denote the corresponding 10th and 90th percentiles. These results were published in [125].

We further analyze the relation between disorder content and the predictive performance in Figure 7. Figure 7A shows average predictive performance of the 13 computational disorder predictors for proteins binned into discrete ranges based on their disorder content. This Figure

shows a clear trend of declining predictive performance with the increasing native disorder content for both metrics: AUC and ACC. The predictive performance stays relatively high until the native disorder content stays below 0.5 and drops substantially when the disorder content exceeds 0.7. The median values of the average protein-level accuracy and AUC equal only to 0.387 and 0.642, respectively, for the proteins which has disorder content above 0.7. The whiskers, which denote the 10th and 90th percentiles, reveal that many proteins with over 0.7 disorder content secure near random levels of accuracy and AUC. Figure 7B illustrates this relation using the previously defined sets of easy-to-predict and hard-to-predict proteins. It compares the distribution of the disorder content between these two proteins sets. It is clear that the easy-to-predict proteins have much lower disorder content in the case of both metrics: AUC and accuracy. These findings are relatively troubling, as they reveal that the current disorder predictors perform relatively poorly for the proteins where they are arguably needed the most (proteins with the most disorder).

3.4 Complementarity and relative protein-level performance of disorder predictors

This section investigates significance of differences in the predictive performance for the 13 predictors and also analyzes complementarity of their predictions. First, we quantify the significance of the differences. The results are shown using arrows in Figure 8. We assume that a given pair of predictors is significantly different if the corresponding p -value < 0.01 . The analysis suggests that majority of the differences are significant, which means that the improvements are consistent across majority of proteins. When it comes to the ROC-AUC, the difference between the best performing SPOT-Disorder and DISOPRED3 is not significant, while they both significantly outperform the other 11 disorder predictors. In terms of accuracy, DISOPRED3 significantly outperforms the other 12 disorder predictors. ESpritz-DisProt that has the lowest median protein-level FPR and has significantly lower FPR values when contrasted with each of the 12 other predictors. The reason behind this low FPR for ESpritz-DisProt is that it under-predicts the amount of disorder. This was also shown in [136], where ESpritz-DisProt predicts 2.6% disorder content in a large dataset with 5% native disorder content, while the other nine methods

considered in that article predict disorder content between 6% (IUPred-long) and 29% (DisEMBL-HL).

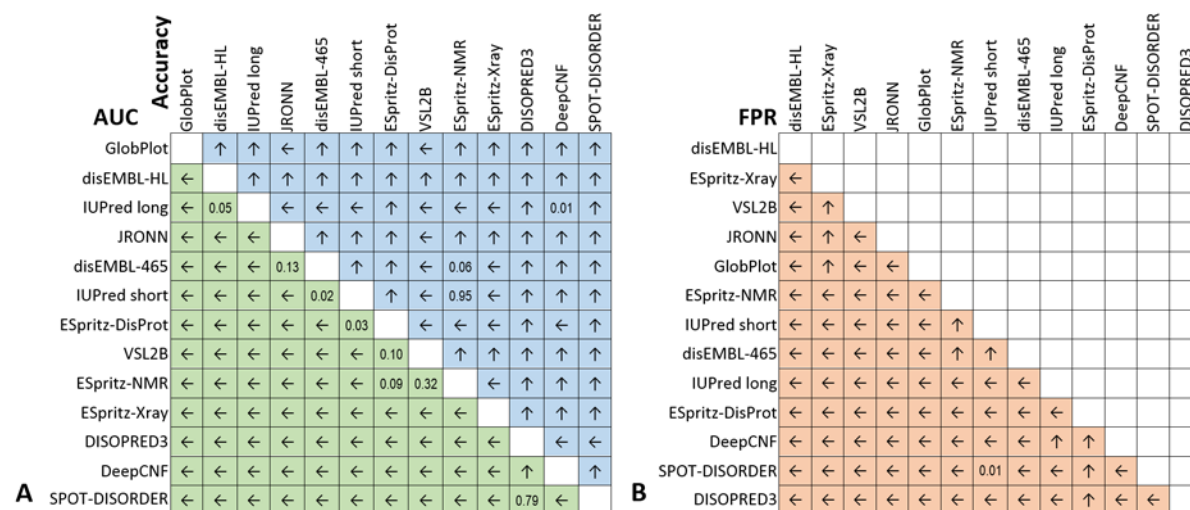


Figure 8: Comparison of the protein-level predictive performance between the 13 disorder predictors. Panel A summarizes comparison of the AUC values (on green background) and accuracies (on blue background). Panel B considers the false positive rates (on red background). Statistical significance of the differences between all pairs of methods was assessed with the t-test for normal measures and otherwise with the Wilcoxon rank-sum test. Normality was tested with the Anderson–Darling test at 0.05 significance. We assume that the difference in predictive performance for a given pair of predictors is significant if the corresponding P-value is < 0.01 . Arrows point to the methods that secure significantly better predictive performance ($P < 0.01$). The P-values are shown for the pairs of methods that are not significantly different. These results were published in [125].

We analyze complementarity of the 13 computational disorder predictors by computing the Pearson’s correlation coefficients (PCCs) for each pairwise comparison of the protein-level AUC-ROC and accuracy values. Figure 9 illustrates the mutual correlation grid for accuracy and ROC-AUC in panel A and panel B, respectively. The color-coding is used to denote different ranges of PCC values where red denotes weak correlation ($PCC < 0.3$); yellow denotes modest correlation ($0.3 > PCC > 0.66$) and green denotes high correlation ($PCC > 0.66$). We find two clear clusters of computational disorder predictors when performing assessment with accuracy (Figure 9A). The low correlation cluster (shown in red) includes six disorder predictors: disEMBL-HL, ESpritz-DisProt, ESpritz-Xray, ESpritz-NMR, GlobPlot and DISOPRED3. The high correlation cluster (shown in green) has three sub cluster groups as follows: i) VSL2B and JRONN; ii) IUPred-long, IUPred-short, disEMBL-465 and DeepCNF; and iii) disEMBL-465, DeepCNF and SPOT-Disorder. Figure 9B

illustrates this analysis for ROC-AUC. We identify three highly correlated (green) clusters: (i) ESpritz-Xray and DeepCNF; (ii) disEMBL-465, JRONN, VSL2B, IUPred-long, IUPred-short and disEMBL-HL; and (iii) SPOT-Disorder and DISOPRED3. Overall, we show that certain methods provide similar levels of predictive performance for the same proteins while they differ substantially from other predictors. This opens an interesting and worth further pursuit opportunity to design meta-predictors that combine results produced by multiple disorder predictors to improve predictive performance.

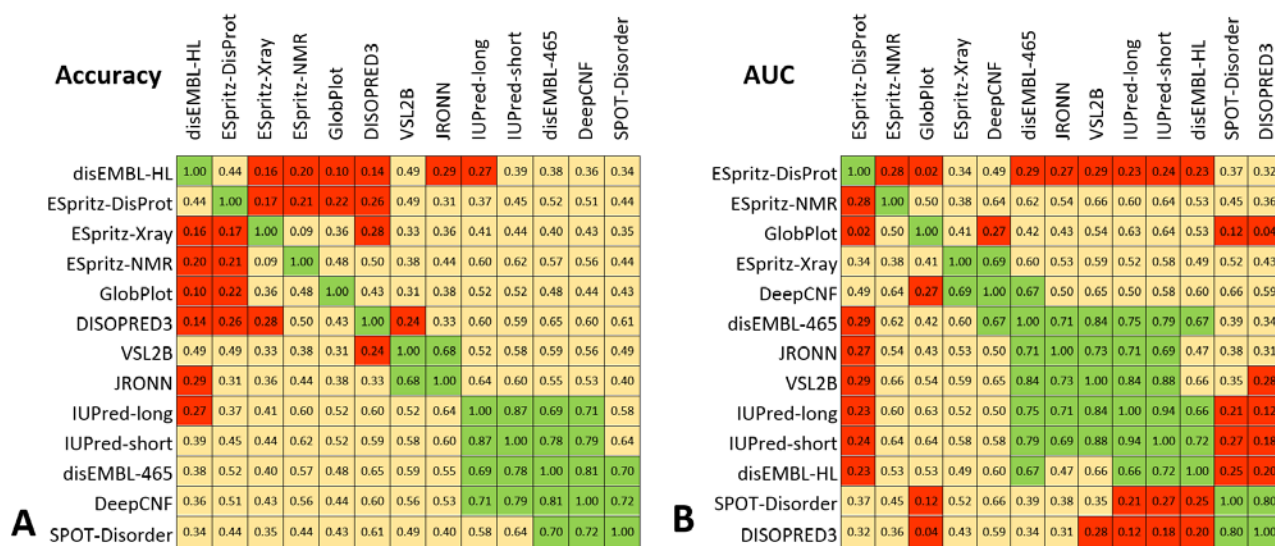


Figure 9: Pearson correlation coefficients (PCCs) between the protein-level predictive performances for each pair of the considered 13 disorder predictors. Panels A and B quantify the performance with accuracy and AUC, respectively. Both correlation matrices are symmetric. The sorting of the predictors differs between the two panel and was optimized to highlight clusters of highly correlated methods. Values of the PCC are color-coded where red denotes no correlation ($PCC < 0.3$), yellow denotes modest correlation ($0.3 \leq PCC \leq 0.66$) and green corresponds to high correlation ($PCC > 0.66$). These results were published in [125].

Along the lines of the complementarity analysis, we assess contributions of individual disorder predictors to produce high quality predictions at the protein level. Figure 10A shows the fraction of proteins for which a given disorder predictor generates the most accurate result, as quantified with the accuracy (inner ring) and AUC (outer ring). The best performing at the dataset-level DISOPRED3 and SPOT-Disorder provide the best protein-level results for only about half of the proteins. Moreover, each of the 13 tools generates the most accurate protein-level predictions for some proteins. This includes the worst dataset-level performers, GlobPlot and disEMBL-HL,

which provide the best accuracies for 0.9% and 0.5% of proteins, respectively, and the highest AUCs for 0.9% and 1.7%, respectively. Figure 10B breaks down the proteins for which a given number of predictors offers highly accurate predictions, i.e. predictive quality that is higher than the expected quality of the best method (the dataset-level performance of the best method). The Figure reveals that only less than 19% of proteins lack highly accurate predictions (dark red regions in Figure 8B). Furthermore, over half of the proteins secure highly accurate predictions (measured with either accuracy or AUC) by at least four disorder predictors, while 25% of proteins (when using accuracy) and 39% of proteins (when using AUC) have such accurate predictions produced by the majority of the 13 disorder predictors. The bottom line is that high-quality protein-level predictions can be often obtained from several disorder predictors. This suggests that the end users should not limit themselves to using only the most accurate (at the dataset-level) methods.

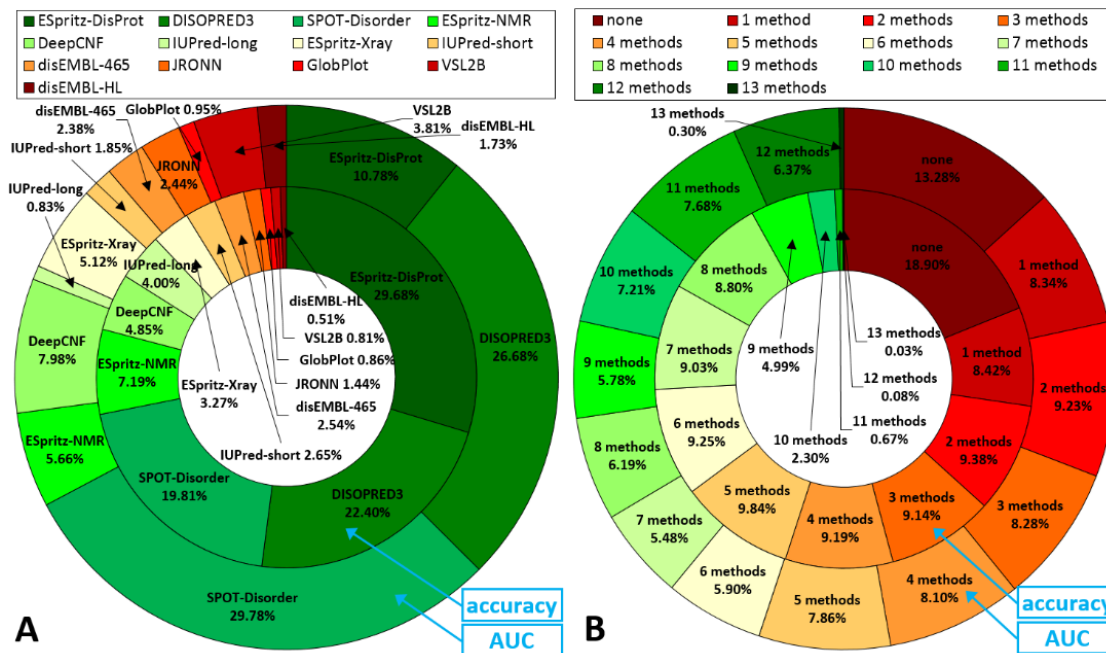


Figure 10: Contributions of the 13 disorder predictors to the production of the highly accurate predictions. Panel A quantifies the fraction of proteins for which a given method generates the highest predictive performance compared to all other disorder predictors. Panel B show the fraction of proteins for which a given number of predictors offer highly accurate predictions, i.e. predictive performance that is higher than the expected performance of the best method (the dataset-level performance of the best method). The inner and outer rings show results when using accuracy and AUC, respectively. These results were published in [125].

3.5 Case Study

This section demonstrates predictions from five selected computational disorder predictors for hydroxymethyltransferase from *Mycobacterium tuberculosis* (Uniprot id: P9WIL7). The native disorder annotation for this protein was obtained from missing electron densities in PDB crystal structures (PDB ID: 1OY0). Figure 11 illustrates predictive propensity scores and binary predictions from DISOPRED3, SPOT-Disorder, VSL2B, IUPred-long and GlobPlot. These five computational disorder predictors are selected to represent three different ranges of the overall predictive performance: DISOPRED3 and SPOT-Disorder for high quality predictions; VSL2B and IUPred-long for medium range and the GlobPlot for the low quality range. The selected five predictors comfortably exceed their dataset-level ROC-AUCs for this protein: DISOPRED3 (AUC = 0.96 for this protein versus 0.90 at the dataset-level), SPOT-Disorder (0.96 versus 0.90), VSL2B (0.88 versus 0.81), IUPred-long (0.93 versus 0.73) and GlobPlot (0.80 versus 0.63). The native disorder annotation for this protein includes three disordered regions (shown in red in Figure 11). While the N-terminus region is was predicted reasonably well by all five disorder predictors, the short IDR at the C terminus is detected by only three out of five predictors. Moreover, only GlobPlot and IUPred-long found the disordered region in the middle of the sequence. The very high AUCs of SPOT-Disorder and DISOPRED3 comes from the fact that these methods produce high propensities for the disordered region at the N-terminus while also predicting low propensities for the structured regions. Moreover, outputs of these two methods show spikes in their predicted propensities near the two other disordered regions. While these spikes are not high enough to trigger generation of the binary disorder prediction (they are below their thresholds shown using the dashed horizontal lines), they suggest that disorder is more likely in these regions than in the other parts of this sequence. Overall, this case study shows how the disorder predictors beat their dataset-level predictive performance, which is a typical scenario that we revealed in this analysis.

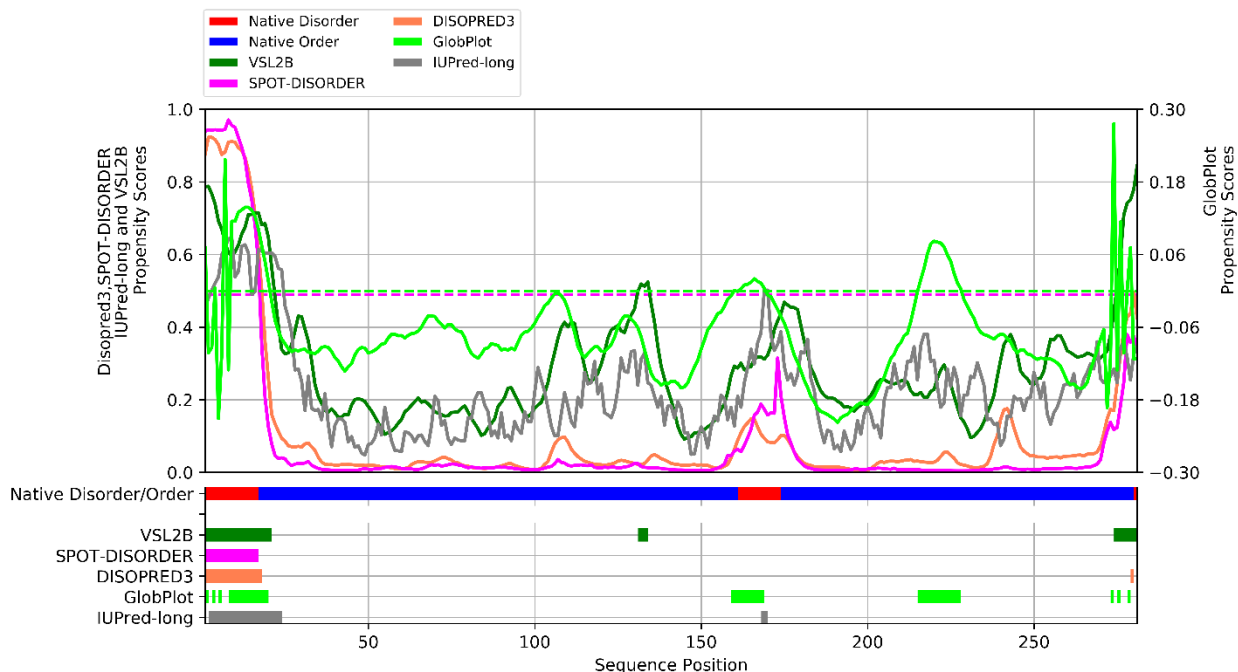


Figure 11: A case study that compares disorder predictions for the hydroxymethyltransferase protein from *M. tuberculosis* (Uniprot id: P9WIL7) that were generated by five methods: VSL2B (dark green), SPOT-DISORDER (magenta), DISOPRED3 (orange), GlobPlot (lime) and IUPred-long (gray). The putative propensities are shown using the solid, color-coded lines. The corresponding binary predictions are given using the color-coded horizontal bars at the bottom of the figure; thresholds that are used to convert the propensities into the binary predictions are visualized with the dashed horizontal lines in the top part of the figure. The red and blue horizontal bar denotes the native annotation of disordered and structured regions, respectively, which were annotated using crystal structure (PDB ID: 1OY0). These results were published in [125].

3.6 Summary

This chapter addresses assessment of the computational disorder predictor from a novel and important viewpoint. Our first-of-its-kind large-scale analysis of 13 representative disorder predictors shows that the quality of the protein-level predictions is often very different from the corresponding dataset-level results. This assessment shows that the protein-level predictive performance is in fact higher than the corresponding dataset-level assessments values for a significant majority of proteins, as of many as over 70% of proteins for the ESpritz-Xray, SPOT-Disorder and DISOPRED3 methods. However, at the same time, we show there are also relatively many poorly predicted proteins for every considered disorder predictor. The fraction of the poor predictions falls within the range of 10% to 30% of the proteins.

Our investigation further reveals that the predictive performance of disorder predictors is correlated with the amount of the native disorder content in a given protein. We observed a trend where performance of disorder predictors drops for proteins with higher disorder content. We demonstrate that easy-to-predict proteins are characterized by low amounts of disorder while the hard-to-predict proteins typically have substantial amounts of disorder. Furthermore, we investigate complementarity of the predictive performance across the 13 disorder predictors as well their relative performance. We show that while two methods, SPOT-Disorder and DISOPRED3, are significantly better than the other predictors, the other methods also provide good quality results for some of the proteins.

Our analysis suggests that disorder predictors provide complementary results. This reveals an interesting opportunity to develop a recommender system that would suggest the most suitable disorder predictor(s) for a given protein sequence. This recommendation system would need to rely on accurate linking of unique characteristics of the input protein sequence, such as physiochemical properties of its amino acids, with the predictive performance of a given predictor. The motivation to pursue this objective comes from the large number of the available disorder predictors, which as we show offer complementary results, and the fact that end users would be undoubtedly overwhelmed by the task of selecting a suitable predictor for their protein of interest.

Chapter 4. Development of a novel protein-level predictor recommendation system to improve predictive performance of disorder predictions

The previous chapter investigated predictive performance of computational disorder predictors. It concludes that the performance cannot be simply represented by the dataset level values and it also must consider results for individual proteins. Furthermore, Chapter 3 has revealed that some of the current disorder predictors perform in a complementary fashion when tested on individual proteins. In other words, some predictors may perform very well while some other may secure poor performance on the same protein, irrespective of their performance on benchmark datasets.

Motivated by the findings from Chapter 3, here we describe the work towards objective 2 that we define in Section 1.3. The results and methods reported in this chapter were published in [140]. This chapter investigates answers to two questions:

1. Is it possible to predict the predictive performance of a given disorder predictor for a given protein?
2. Given that answer to question one is positive, could this prediction be used to identify well-performing disorder predictors for a given protein?

We cover a comprehensive set of 12 disorder predictors representing the three categories of methods that were described in Section 2.2.1. These computational disorder predictors are selected based on their relatively low runtime, availability and previously reported predictive performance. The estimation of the predictive performance for given proteins is based solely on the physiochemical properties of the input amino acid sequence and hence the only input that we use is the readily available sequence. The overarching goal is to suggest the disorder predictor that provides the highest performance for a given input protein sequence. As a byproduct of our approach, we will also offer an estimate of the expected predictive performance that will be

produced by the selected predictor, giving the end users useful information whether to pursue the prediction and suggesting which particular predictive tool to use.

4.1 Protein-level predictive performance of disorder predictors

We use the AUC-ROC as the performance evaluation measure. This is the most widely used metric to evaluate the predictive performance of disorder predictors [13, 15, 18, 79, 128].

This section evaluates per protein ROC-AUC for 12 representative computational disorder predictors that include three versions of ESpritz that predict intrinsic disorder annotated from x-ray structures (ESpritz-Xray), NMR structures (ESpritz-NMR) and using DisProt database (ESpritz-DisProt)[59]; the two versions of IUPred that are optimized to predict disorder short regions and disorder long regions (IUPred-short and IUPred-long, respectively)[48]; the two versions of DisEMBL which cover X-ray structures (DisEMBL-465) and loop secondary structures (DisEMBL-HL) [50]; GlobPlot [47], RONN [49], VSL2B [53], DISOPRED3 [100] and SPOT-Disorder [100]. The evaluation was done on the dataset of 5,272 proteins that was described in the Section 3.2. Since calculation of AUC requires presence of both disordered (positive) and ordered (negative) residues, we limit this analysis to proteins that have at least four ordered and at least four disordered residues. Figure 12 illustrates distribution of ROC-AUC values for the corresponding set of 3,621 proteins and the 12 selected predictors.

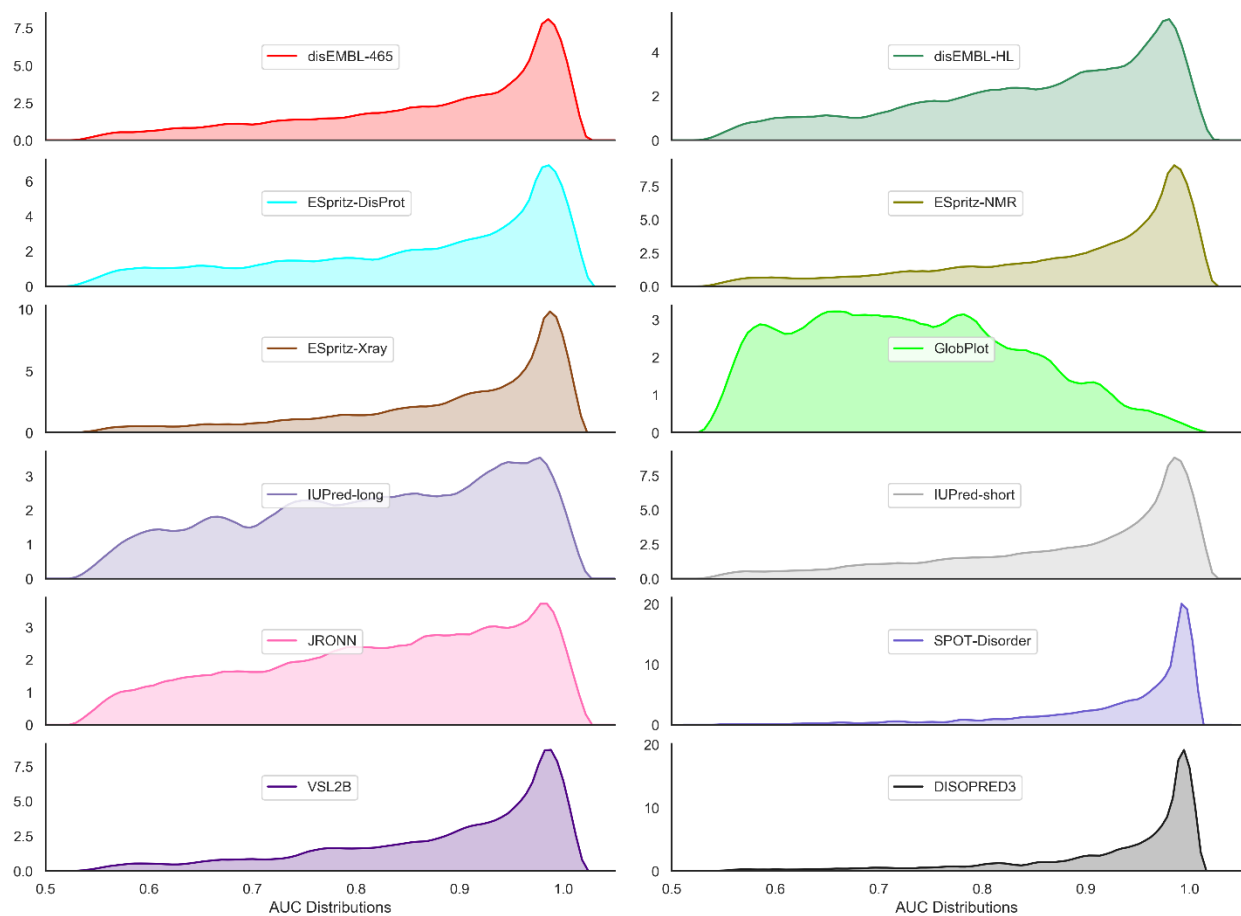


Figure 12: Distribution of per-protein AUC values 12 computational disorder predictors over 3,126 proteins in the benchmark dataset. These results were published in [140].

Virtually all considered disorder predictors, except for GlobPlot, have long tails of poorly predicted proteins and the bulk of the predictions characterized by high ROC-AUC values, which agrees with results from Chapter 3. This shows that this pattern of predictive performance is not affected by the above selection of the proteins that have both disordered and structured residues. The DISOPRED3 predictor has the highest reported dataset level ROC-AUC of 0.9 [17] and it secures a sharp peak with relatively low but still fairly long tail of poor-quality predictions. As another example, VSL2B has lower dataset level ROC-AUC of 0.82 [126] and we observe a wider peak with a bigger tail when compared to DISOPRED3. As we noted in Chapter 3, AUC values for individual proteins vary widely from the dataset-level values. For instance, for VSL2B that has the dataset-level AUC = 0.82, Figure 12 reveals that 59.2% proteins have AUCs >0.9 while 9.3% of proteins are predicted with AUCs <0.6. Similarly, for DISOPRED3 that has the dataset-

level AUC = 0.90, 9.5% of proteins are predicted with AUCs <0.8 while 68.5% have AUCs >0.95. This wide range of the per-protein predictive performance motivates the development of a tool that estimates performance for a given protein.

4.2 Experimental workflow

We aim to develop a system that recommends a well-performing predictor for a given protein sequence. The first step is to predict the expected ROC-AUC for a given protein and a given disorder predictor. Next, we simply recommend the predictors with the highest predicted performance. This system offers practical value in two novel ways. The first is to assist the users in selection of a good disorder predictors for a specific protein. This is valuable even for experienced end users who are well-informed about the availability and quality of different disorder predictors. This is because the benchmark results for disorder predictor may vary widely from their protein-level performance. The second is to offer an estimate of the performance that will be produced by the selected predictor. This will inform the user to which extent they can trust the prediction and will allow them to judge whether to pursue the prediction in the first place.

4.2.1 Architecture and design of the protein-level predictors of disorder prediction quality

This first step provides estimate of the expected ROC-AUC values of a disorder predictor for a given input protein. These estimates are produced by machine learning models. Figure 13 illustrates the architecture of the proposed method. Panel (a) gives the flowchart while panel (b) provides a pseudocode for this prediction process.

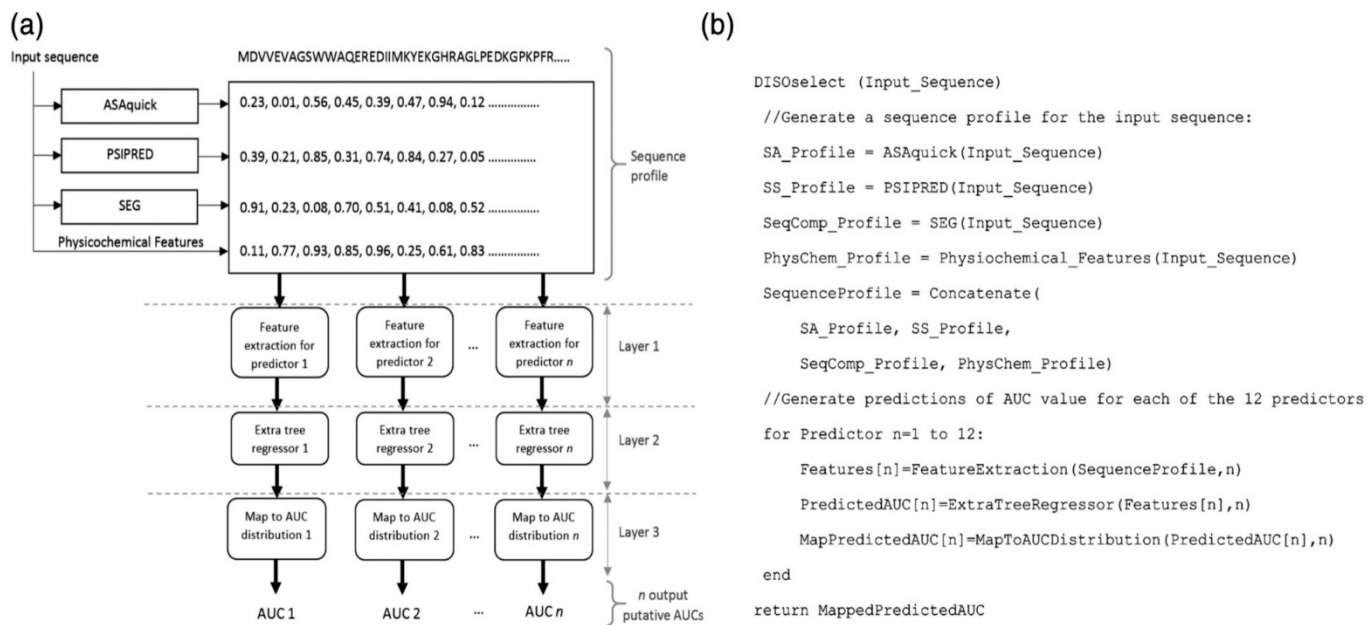


Figure 13: Architecture of the proposed recommendation system. Panel (a) gives flowchart of the proposed model while panel (b) shows the corresponding pseudocode. This illustration was published in [140].

The prediction begins by collecting the amino acid sequence of the proteins as the sole input. Next, we use the amino acid sequence to generate a sequence profile that provides additional residue-level information about the input protein. This sequence profiles consist of a few physicochemical and structural properties that are related with presence of disorder [25, 32, 48, 141-145]. They include two putative structural properties generated by third-party predictors: solvent accessibility predicted with ASAquick [146] and secondary structure predicted with the single sequence version of the PSIPRED [147]. The selection of these structural predictors was based on their availability and computational efficiency. The profile also includes sequence complexity generated by the SEG algorithm [137] and a few sequence-derived physicochemical properties like hydrophobicity, hydrophathy, charge, structural entropy, polarity, volume, size, flexibility, refractivity, transfer and solvation energies, and propensity for coil, turn, strands, helix and disordered conformations. These physicochemical properties rely on the corresponding indices that we collected from AAindex database [148]. We encode the sequence profile into a vector of 130 numeric features that aggregate the information concerning structural and physiochemical properties at the whole-protein level. These features are detailed in the Appendix 1. They include 21 features computed directly from the input sequence (AA composition and sequence length), 3 features computed from the putative solvent accessibility, 2 features from the sequence complexity, 8 features from the putative secondary structure, and 96 features based on the physiochemical properties.

Figure 13 summarizes the architecture of our predictive models using the 3-layer design. The layers 1 and 2 refer to the generation of the feature vector from the sequence profile which is processed by the predictive model, in the form of a regression tree. We performed feature selection and optimization of the predictive model using the training set of 5,272 proteins described in Section 3.2. From the machine learning perspective, this task is a regression problem where each machine learning model is trained to predict the expected ROC-AUC for a given disorder predictor.

The feature selection was done in two steps. First, features with high mutual correlation were removed in order to filter out similar features. We quantified the mutual similarity for all pairs of the 130 features based on the Pearson correlation coefficients (PCCs). For each pair of highly

correlated features ($|PCC| > 0.65$), one of them that has lower predictive performance was removed. The predictive performance was measured based on the PCC with the per-protein AUC scores for a given disorder predictor. This step resulted in the removal of between 21 and 40 features, depending on the predictor.

Next, we removed features that have low predictive power via wrapper-based feature selection, done separately for each of the 12 disorder predictors. In short, we screened features by their ROC-AUC values when used individually to predict the protein-level AUCs of a given disorder predictor. A subset of features with $|PCC| > \text{threshold}$ was selected, where the value of the threshold was selected to provide the highest predictive quality assessed by the three fold cross validation on the training set. The threshold during was gradually increased by 0.01 starting from 0 until the cross-validated PCC between actual ROC-AUC and predicted ROC-AUC drops. The selected feature sets ranged in size between 24 and 38, depending on the disorder predictor used.

As part of the wrapper-based feature selection and optimization of the predictive model, we considered three types of regression models: nearest neighbor regression [149], linear regression [150], and extra tree regression [151]. The corresponding average (over different disorder predictors) PCCs on the training set were 0.15, 0.13 and 0.31, respectively. The average mean squared error (MSE) values for the same three algorithms were 0.018, 0.011 and 0.007, respectively. Hence, the extra tree regression was selected the best performing predictive model. The extra tree regressor is an enhanced version of extremely randomized random forests that relies on randomization to grow trees [152]. The key objective behind this randomization is to minimize overfitting into the training set. The avoidance of overfitting is specifically important in our case as the training set and test set share low (<25%) similarity. Furthermore, the extra tree regression is more computationally efficient than the conventional random forests. We parametrized our regression models via a grid search for each of the 12 disorder predictors, with the aim to maximize the PCC for the 3-fold cross validation on the training dataset. For instance, for the extra tree regression we parameterized the maximum depth of the trees (using 0–20 range), number of trees in the forest (100–180 range), minimum number of samples required for

a split (0–30 range), and the number of features to consider when calculating best split (0–50 range).

4.2.2 Design of the protein-level disorder predictor recommendation system

In the second step the 12 extra tree regression-generated estimates of ROC-AUCs are processed and compared and the disorder predictor with the highest expected ROC-AUC is recommended to the user. More specifically, predictions from each of the 12 regressors are mapped into the distribution of AUCs values from the training datasets for the corresponding disorder predictor (third layer in Fig. 13). This ensures that the predicted AUCs are calibrated to cover the entire spectrum of AUC values that are produced by a specific disorder predictor. Finally, the 12 predicted AUCs are compared and the method with the highest putative AUCs is recommended back to the user. We call the complete recommendation system DISOselect. The outputs generated by DISOselect include the recommended disorder predictor, its putative AUC and the AUC values for each of the other 11 predictors.

4.2.3 Analysis of the predictive model

This section investigates how our model accomplishes the prediction of the expected AUCs values. More specifically, we analyze relation between the input features and the output (AUC values). We investigate individual features produced in the layer 1 of the model (Figure 13) and the importance of specific features groups to the predictive performance of the extra tree regressor.

In order to analyze the contribution of individual features, we focus on the top two features with the highest PCC with the output for each of the 12 considered disorder predictor. After selecting top two most correlated features, we obtained 18 features since some of these features were shared by multiple disorder predictors. The predictive value of these features is summarized in Figure 14 where darker shading corresponds to a higher magnitude of correlation with the output (stronger contribution to the model). The direction of the arrows denotes the sign of the correlation where upwards pointing arrows are for positive correlation and downwards pointing arrows are for negative correlations.

As one highlight, we observe that the average accessible surface (ASA) area is a strong contributor to our model (bottom of Figure 14). Interestingly, ASA is positively correlated with the SPOT-Disorder’s and DISOPRED3’s AUCs, but negatively correlated with the ESpritz-DisProt’s performance. In general, we note that virtually all features switch the sign of the correlation when used to make predictions for different disorder predictors. This demonstrates each disorder predictor has its own unique predictive pattern. Individual performance biases may reflect the makeup of the training sets and the selection of the sequence-derived predictive inputs utilized by individual predictors.

		PREDICTIVE MODELS											
		disEMBL-465	IUPred-short	disEMBL-HL	ESpritz-DisProt	ESpritz-NMR	ESpritz-Xray	GlobPlot	IUPred-long	JRONN	VSL2B	SPOT-Disorder	DISOPRED3
TOP SELECTED FEATURES	Count of strands	↗	↗	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘
	Summed native hydrophobicity (CASG920101)	↗	↗	↗	↘	↘	↗	↗	↗	↘	↗	↗	↗
	Count of coils and strands	↘	↗	↗	↗	↘	↘	↘	↘	x	↗	↘	x
	Summed beta sheet frequency (PALJ810104)	↘	↘	↗	↗	↘	↘	↘	↘	x	↘	↘	↘
	Summed average surrounding hydrophobicity (MANP780101)	↘	↘	↘	↗	↘	↘	↘	x	↗	↘	↘	↘
	Summed normalized flexibility (VINM940103)	↘	↘	↘	↗	↘	↘	↘	x	↗	↘	↘	x
	Summed charge transfer (CHAM830107)	↘	↘	↘	↗	↗	↘	↘	x	↗	↘	↘	↘
	Count of coils	↘	↘	↘	↗	↗	↘	↘	x	↗	↘	↘	↘
	Summed refractivity (MCMT640101)	↘	↘	↘	↗	↘	↗	↘	x	x	↘	↘	↘
	Content of coils and strands	↘	↘	↘	↗	↘	↘	↗	↘	↘	↘	↘	↘
	Content of helices	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
	Average normalized flexibility (VINM940103)	↗	↗	↘	↘	↗	x	↗	↗	↗	↗	↗	↘
	Average consensus normalized hydrophobicity (EISD840101)	↘	↘	↗	↗	↘	↘	↘	↗	↘	↘	↗	x
	Average buried fractions (JANJ790101)	x	↘	x	↗	x	↘	↘	↘	↗	↗	x	x
	Average beta sheet frequency (PALJ810104)	↘	↘	↘	↗	↘	x	↘	↘	↗	↘	↘	↘
	Average native hydrophobicity (CASG920101)	↘	↘	↗	↗	↘	↘	↘	↘	↘	↗	↗	↗
	Average NH chemical shifts (BUNA790101)	↘	x	x	↘	↘	↘	↗	↘	↘	↘	↗	↘
	Average accessible surface area	↗	↗	↗	↘	↗	↗	↗	↗	x	↗	↗	↗

Figure 14: Key predictive features used to predict AUC of the 12 disorder predictors. The predictive performance of individual features is quantified with the Pearson correlation coefficients (PCC) between feature values (horizontal lines) and the prediction output (actual AUC) for each disorder predictor (vertical lines) that were quantified on the training dataset. Detailed explanation of features is available in Appendix 1. PCC values are color-coded where dark green is for $|PCC| \geq 0.3$, light green for $|PCC|$ between 0.15 and 0.30, white for $|PCC| < 0.15$, and grey with ‘x’ symbol indicate the a given feature is not included in the model for that predictor. The direction of arrows reveals the sign of PCC where upwards arrows denote positive correlation while downward arrows denote negative correlation. These results were published in [140].

We also analyze importance of the five feature categories that are described in Section 4.2.1 and which are defined by the data in the sequence profile that was used to produce them. The importance was quantified with information gain, which measures decrease in the classification entropy due to the use of a given feature. This is motivated by the use of the entropy in extra-tree regressor models [153]. The results for each of the considered 12 disorder predictors are summarized in Figure 15. The dominant feature category (features extracted from the putative secondary structure) is consistent for all 12 computational predictors, although the magnitude of the contribution varies considerably. The predictive value of the secondary structure can be explained by its relation to the intrinsic disorder. The disorder is poorly predicted in the proteins with higher fractions of the secondary structures (helices and strands) given that these proteins are primarily structured. On the other hand, disorder in the proteins heavily composed of coils is in general well predicted. Moreover, putative secondary structure is also used as the predictive input by some of the considered here disorder predictors, such as MFDp2 [129], CSpritz [64] and Spritz [154]. The second and third-best categories of features are the physiochemical properties and AA composition. They are closely related as both are extracted directly from the protein sequence. This reveals an implicit bias of the disorder predictors for which performance depends on the amino acid composition and physiochemical characteristics of the protein. The two least-contributing feature categories are the sequence complexity and solvent accessibility.

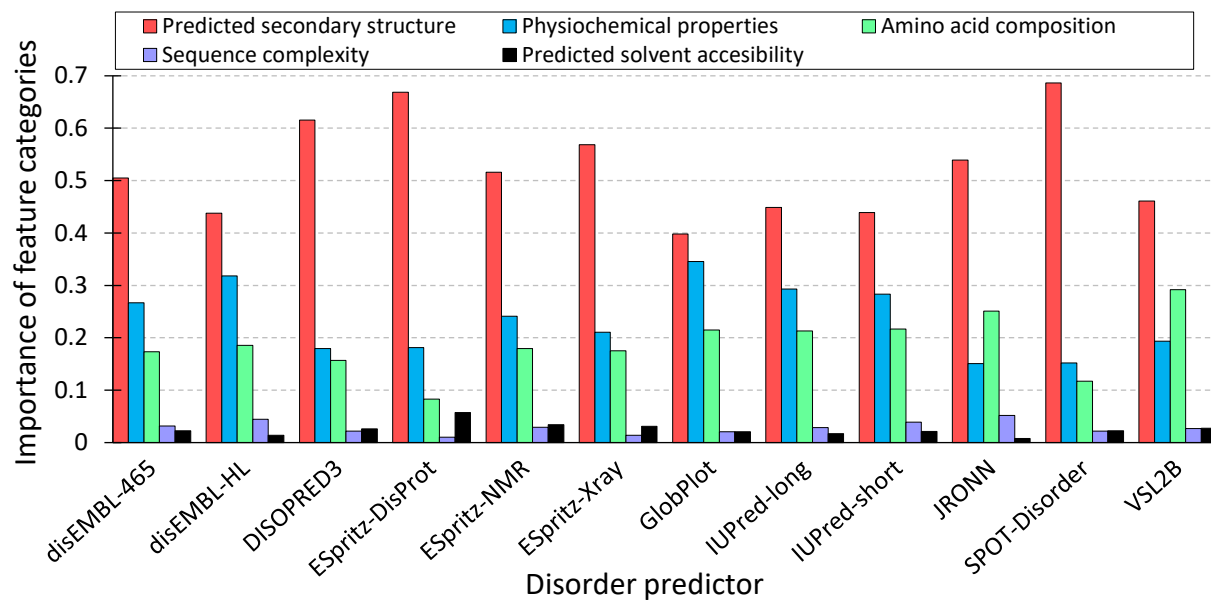


Figure 15: Importance the five feature categories for the predictive models designed for the 12 disorder predictors. We used a three-step process to derive the scores for each predictive model. First, the information gain of individual features was calculated from the extra-tree regressors. Second, features were divided into the five classes and the information gain of the features in the same category was summed up. Third, the summed values were divided by the sum of the information gain values of all features in the same model. The last step allow for directly comparison of relative contributions of each feature category. These results were published in [140].

To summarize, the ability of the regressors to predict AUCs of the 12 disorder predictors primarily depends on the information extracted from the predicted secondary structure, physicochemical properties and amino acid composition of the input protein sequence. Furthermore, the predictive quality of specific disorder predictors is governed by unique set of relations, which is why it is necessary to build and optimize regressors individually for each disorder predictor.

4.3 Assessment of the protein-level disorder predictor recommendation system

4.3.1 Predictive performance of the extra tree regressor models

The regressors that were trained and optimized on the training set of 5,272 proteins were next comparatively evaluated on the dissimilar test set of 999 test proteins. We note that the training and test proteins share <25% sequence similarity. The predictive performance of our models is compared to two controls. We note that this is the first-of-its-kind model that predicts

performance of disorder predictors and so we are unable to compare it with the prior published solutions. The first control is a predictor that selects at random an AUC value from the training dataset. This ensures that this predictor generated the correct distribution of the AUC values for a given disorder predictor. The second control is based on sequence alignment/similarity. For a given test protein we use the AUC of the most similar training protein that is selected using the popular BLAST algorithm with default parameters [155, 156].

The predictive performance is quantified with two metrics: (a) mean squared error (MSE) of predicted AUCs (when compared with the native/true AUC) and (b) correlation between the actual ROC-AUC and the predicted ROC-AUC. We use the better of the two controls to calculate the ratio of improvement between our model and controls.

Table 6: Predictive performance of the extra tree regressor-based model and two controls. Mean squared error (MSE) and Pearson correlation coefficients (PCC) values are calculated between the predicted AUC and the actual AUC for each test proteins. Controls were produced using a random and a sequence similarity-based approaches. Paired significance tests were performed between the predicted AUCs of our regressor and the results produced by the controls: [+] denotes that our model is significantly better with p-value <.05. We used the paired t test for normal measurements and the Wilcoxon test otherwise; normality was tested with the Anderson-Darling test at the .05 significance. These results were published in [140].

Disorder predictor	MSE (mean squared error)				PCC (Pearson correlation coefficient)			
	Extra tree regressor	Random control	Similarity-based control	Improvement ratio	Extra tree regressor	Random control	Similarity-based control	Improvement ratio
disEMBL-HL	0.009	0.05 [+]	0.05 [+]	5.6	0.36	0.01 [+]	0.13 [+]	2.8
IUPred-long	0.011	0.05 [+]	0.04 [+]	3.6	0.32	-0.01 [+]	0.18 [+]	1.8
IUPred-short	0.011	0.05 [+]	0.05 [+]	4.6	0.32	0.03 [+]	0.07 [+]	4.6
VSL2B	0.008	0.05 [+]	0.05 [+]	6.3	0.31	0.03 [+]	0.14 [+]	2.2
disEMBL-465	0.008	0.05 [+]	0.05 [+]	6.3	0.30	-0.03 [+]	0.14 [+]	2.1
GlobPlot	0.011	0.05 [+]	0.05 [+]	4.6	0.28	-0.13 [+]	0.09 [+]	3.1
ESpritz-NMR	0.009	0.05 [+]	0.05 [+]	5.6	0.28	0.02 [+]	0.08 [+]	3.5
SPOT-Disorder	0.010	0.05 [+]	0.04 [+]	4.0	0.24	0.00 [+]	0.00 [+]	48.0
DISOPRED3	0.004	0.04 [+]	0.05 [+]	12.5	0.24	-0.11 [+]	0.15 [+]	1.6
ESpritz-Xray	0.007	0.05 [+]	0.05 [+]	7.1	0.23	-0.08 [+]	0.03 [+]	7.7
JRONN	0.008	0.05 [+]	0.06 [+]	7.5	0.19	-0.01 [+]	0.00 [+]	190.0
ESpritz-DisProt	0.007	0.05 [+]	0.04 [+]	5.7	0.12	0.00 [+]	0.05 [+]	2.4

Table 6 summarizes the results and reveals that our extra tree regressor model significantly outperforms both controls in terms of both MSE and PCC (p-value < .05). While the MSE and PCC

values of our predictor are modest (PCC at around 0.3 and MSE at about 0.01), we note that the rate of improvement over the controls is relatively high. The MSE values are between 3.6 and 12.5 folds better than the best control, while correlation values are between 1.6 and 190 folds better than the controls.

4.3.2 Use of the extra tree regressors for the selection of well-predicted proteins

The extra tree regressors aim to identify proteins that can be predicted with higher predictive performance for each of the 12 disorder predictors. This will allow the end users and designers of the future predictors to identify poorly vs. well predicted proteins for a given disorder predictor. Figure 16 investigates whether our regressors can be used for this purpose. First, for each disorder predictor, we sort the test proteins in the ascending order of their regressor-predicted AUCs. Then we progressively remove the 5% of proteins with the lowest predicted AUCs and we calculate the actual AUCs of these protein sets. Figure 16 plots the relation between the predicted and the actual AUC values of these proteins sets.

The raising trends in Figure 16 demonstrate that the selection done based on the predicted AUC aligns with the actual predictive performance. The upwards trends are consistent across all 12 disorder predictors. The differences in the actual AUC values between the results on the entire test dataset (left-most points in Figure 16) and the smallest set of the 5% of proteins with the highest estimated AUCs are very substantial. As an example, for DISOPRED3, the 5% of proteins with the best estimated AUCs secure AUC = 0.950 when compared to AUC = 0.918 on the test datasets, which translates to $(0.950-0.918)/(1-0.918) = 39\%$ error reduction. The largest absolute increase in AUC is for disEMBL-HL where the 5% of the best predicted proteins secure AUC = 0.896 compared to the AUC = 0.761 on the whole test dataset, which corresponds to $(0.896-0.761)/(1-0.761) = 56\%$ error reduction.

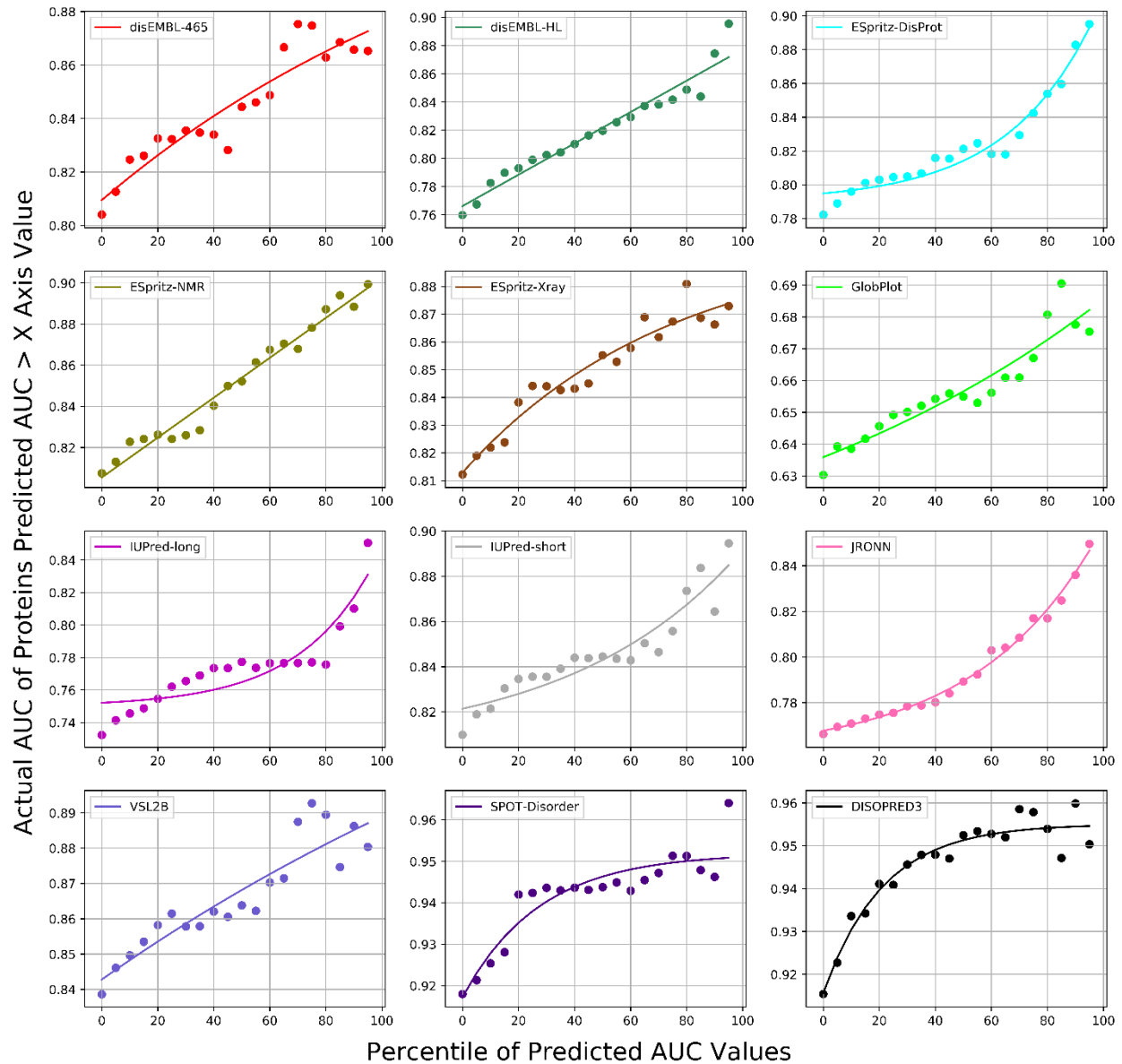


Figure 16: The dataset-level actual AUC values for subsets of the test proteins that are sorted based on their AUCs values estimated by the extra tree regressors. Individual panels correspond to different disorder predictors. Points in each panel correspond to AUCs of the subsets of test proteins for which the estimated AUCs are above a given percentile of all estimated AUCs, that is, the 20 mark on the x-axis corresponds to the 80% of the test proteins that have estimated AUCs that are above the 20th percentile of estimated AUCs generated by the extra tree regressors. The left-most point corresponds to the result on the complete test dataset while the right-most point corresponds to the 5% of test proteins with the highest estimated AUCs. The line is the third-degree polynomial fit into the measured data. These results were published in [140].

Figure 17 summarizes the differences in the actual predictive performance measured using several metrics across the 12 disorder predictors. It shows PR-AUC, sensitivity, MCC and accuracy in addition to the ROC-AUC. The boxes in the Figure 17 represent the distribution of improvement for the given evaluation criteria over the 12 disorder predictors. The improvements are measured by comparing the results on the whole dataset with the results for the three quartiles of selected proteins according to the predicted AUC (75%, 50% and 25%). Positive values indicate improvements. The Figure demonstrates that the improvements are always positive, substantial and consistent for all evaluation criteria. Overall, we conclude that our regressors can be used to accurately predict the well performing proteins for a given disorder predictor.

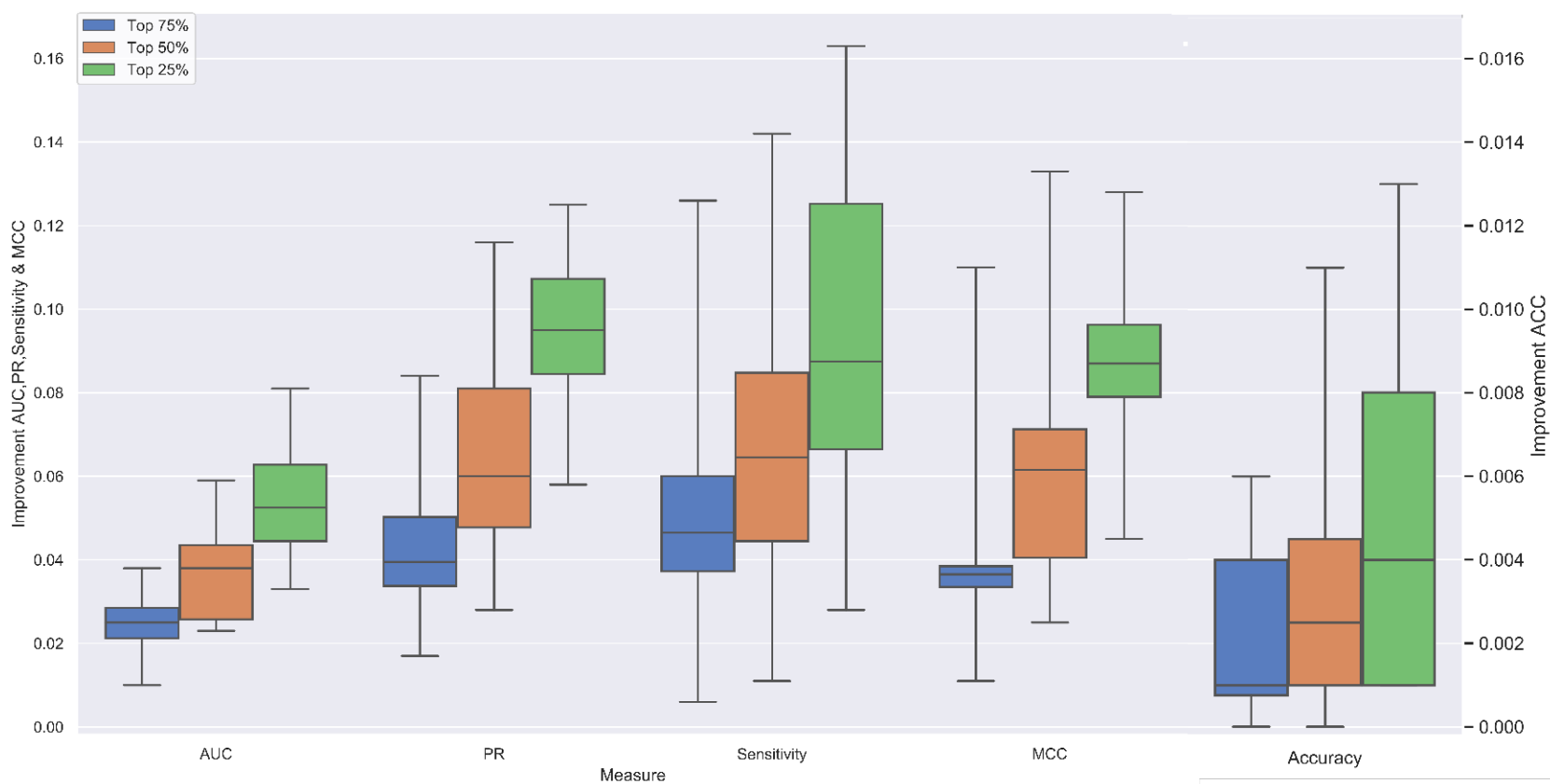


Figure 17: Improvements in the actual ROC-AUC, PR-AUC, sensitivity, MCC and accuracy values computed as the difference between the values for subsets of the top 25% (in green), 50% (in orange) and 75% (in blue) of the test proteins selected based on their AUCs values estimated by extra tree regressors and the whole dataset-level AUCs. Positive values of the improvement indicate that AUC for the subsets of the test proteins are higher than for the complete test dataset. The box plots represent the distribution of the improvements across the 12 disorder predictors where whiskers corresponding to the minimal and maximal improvements and boxes denote the first, second and third quartiles. These results were published in [140].

4.3.3 Predictive quality of DISOselect

This section investigates the performance of the complete recommendation system, DISOselect. We compare the results produced by the predictors recommended by DISOselect with the results of the original disorder predictors used individually. We also compare with conventional meta approaches to combining disorder predictions, where several prediction methods are combined to give a single prediction, usually at the residue level. We consider several residue-level meta-prediction methods based on the 12 individual disorder predictors examined in this work. We use several variations on the meta-predictor construction: two-different architectures – logistic regression (LR) and support-vector regression (SVR), and different input predictors – either all 12 predictors or only the best of the predictors. The best predictors were selected based on the dataset level performance on the training set. Based on these assessments, we selected two prediction methods – SPOT Disorder and Disopred3 – as significantly better than the other individual methods. This gave four meta-predictors: Twelve Predictor LR, Twelve Predictor SVR, Top Two Predictor LR, and Top Two Predictor SVR. The logistic regression model was trained with the 3-fold cross validation on the training dataset with default L2 regularization penalty by balanced class weights according to the proportions of training set using the L-BFGS optimization algorithm. The SVR models were trained with the 3-fold cross validation on the training dataset after subsampling 10% of each fold randomly to minimize the training time. We used the radial basis function kernel and performed a grid search for the penalty parameter C (2^{-5} to 2^5), kernel coefficient γ (0 to 1), and tolerance for stopping criteria (10^{-3} to 10^3).

Figure 18 illustrates the protein level ROC-AUC values of DISOselect on the test set (in dark black thick line) against other 12 individual computational disorder predictors and the two best conventional meta-predictors (thick blue line and thick yellow line). The dark red line denotes a hypothetical oracle predictor which always picks the best performing disorder predictor for each given individual protein. Figure 18 shows that DISOselect performs substantially better than all individual disorder predictors and the conventional consensus approaches. Moreover, the DISOselect's performance is relatively close to the hypothetical best possible approach (oracle predictor).

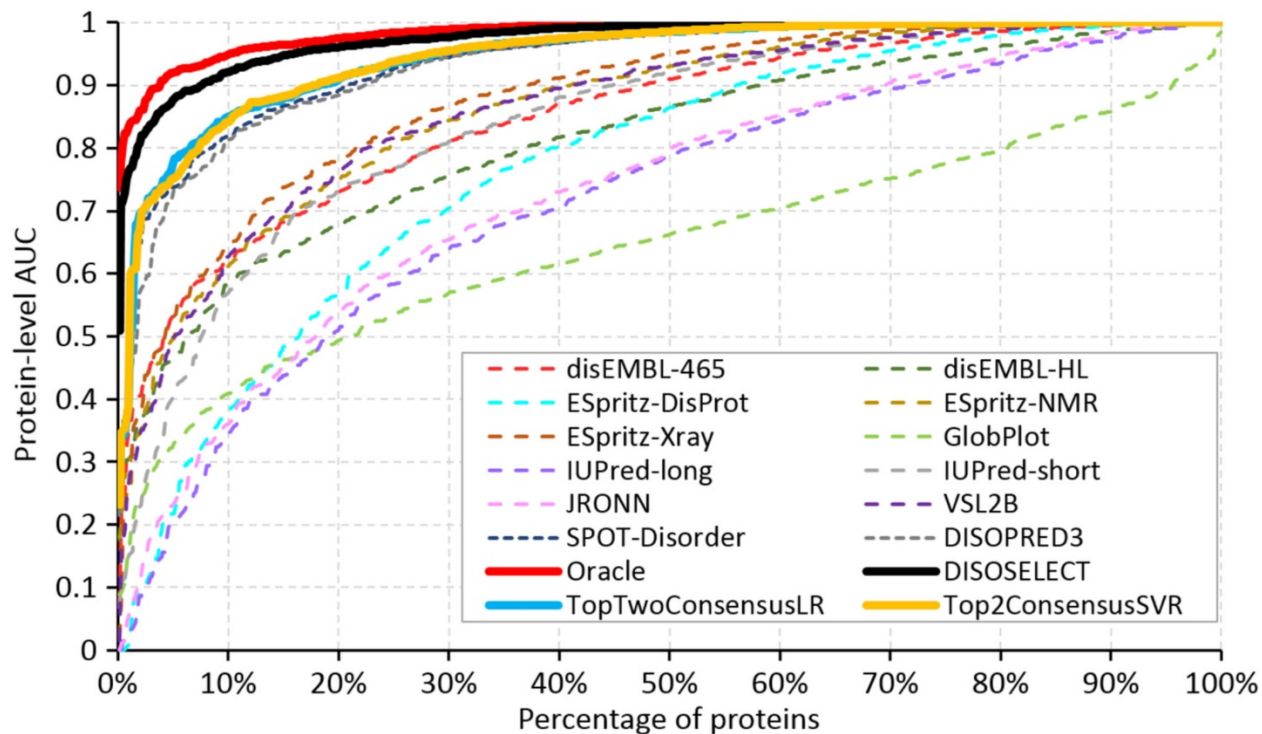


Figure 18: Comparison of the per-protein AUC values between the 12 disorder predictors, the selection of the best disorder predictor using the highest putative AUCs generated by DISOselect (thick black line), and the oracle method (thick red line), and two conventional meta predictors (thick yellow and blue lines) on the test proteins. The oracle method selects the disorder predictor with the highest AUC among the 12 disorder predictors. Lines show the per-protein AUCs that are sorted in the ascending order for each of the considered methods. These results were published in [140].

Table 7 directly compares the performance of the considered models. It compares the mean of per protein ROC-AUC values produced by DISOselect, the twelve individual disorder predictors, and the oracle approach. Statistical tests of significance reveal that DISOselect generates the per-protein ROC-AUC values that are significantly better than the results offered by each of the 12 disorder predictors (p -value < 0.01). The oracle predictor gives the per protein mean ROC-AUC of 0.98 while DISOselect has ROC-AUC = 0.97.

Table 7: Comparison of the per-protein AUC values produced by the 12 disorder predictors, the oracle method that selects the predictor with the highest AUC and the selection based on the highest putative AUC produced by DISOselect for the test proteins. We compared the mean per-protein AUCs computed over the test proteins and the AUCs for the worst (the least accurately predicted) quartile of the test proteins (i.e., the 25% point in Figure 17). Methods are sorted by their mean per-protein AUCs. Significance of the differences in the per-protein AUCs of the predictions selected by DISOselect and the predictions generated by the other methods (including the oracle) was assessed with the *t*-test for normal measurements and the Wilcoxon test otherwise; normality was tested with the Anderson-Darling test at 0.05 significance; we sampled 50% of proteins in the test dataset ten times at random and compared the corresponding 10 pairs of AUCs; the resulting *p*-values are listed in the last column. These results were published in [140].

Predictor	Mean per-protein AUC	Per-protein AUC at the worst quartile of proteins	Significance of differences compared to DISOselect
Oracle	0.983	0.984	<i>p</i> -value<0.01 (significantly better)
DISOselect	0.974	0.971	
SPOT-Disorder	0.940	0.927	<i>p</i> -value<0.01 (significantly worse)
DISOPRED3	0.935	0.921	<i>p</i> -value<0.01 (significantly worse)
ESpritz-Xray	0.880	0.832	<i>p</i> -value<0.01 (significantly worse)
ESpritz-NMR	0.865	0.809	<i>p</i> -value<0.01 (significantly worse)
VSL2B	0.864	0.816	<i>p</i> -value<0.01 (significantly worse)
disEMBL-465	0.853	0.768	<i>p</i> -value<0.01 (significantly worse)
IUPred-short	0.843	0.768	<i>p</i> -value<0.01 (significantly worse)
disEMBL-HL	0.816	0.719	<i>p</i> -value<0.01 (significantly worse)
ESpritz-DisProt	0.772	0.649	<i>p</i> -value<0.01 (significantly worse)
JRONN	0.733	0.603	<i>p</i> -value<0.01 (significantly worse)
IUPred-long	0.718	0.584	<i>p</i> -value<0.01 (significantly worse)
GlobPlot	0.646	0.537	<i>p</i> -value<0.01 (significantly worse)

The analysis in Figure 18 is aggregated at the test dataset level for clarity, i.e., we compared re-sorted per-protein AUCs across different methods. Figure 19 offers a direct comparison of predictive performance of the results selected with DISOselect against the most accurate individual disorder predictor (SPOT-Disorder), the best performing meta-prediction method (Top Two Predictor SVR), and an average disorder predictor. When compared against SPOT-Disorder (red line), DISOselect selects a better disorder prediction for 64% of proteins, the same prediction for 5% of proteins, and worse prediction for 31% of proteins and has average overall improvement of 0.035. DISOselect is better for 64% of proteins, equal for 4% of proteins and performs worse for 34% proteins when compared to the best conventional consensus method (green line). Finally, DISOselect performs better than average of the 12 computational disorder predictors for 95% of proteins with overall average improvement of 0.152.

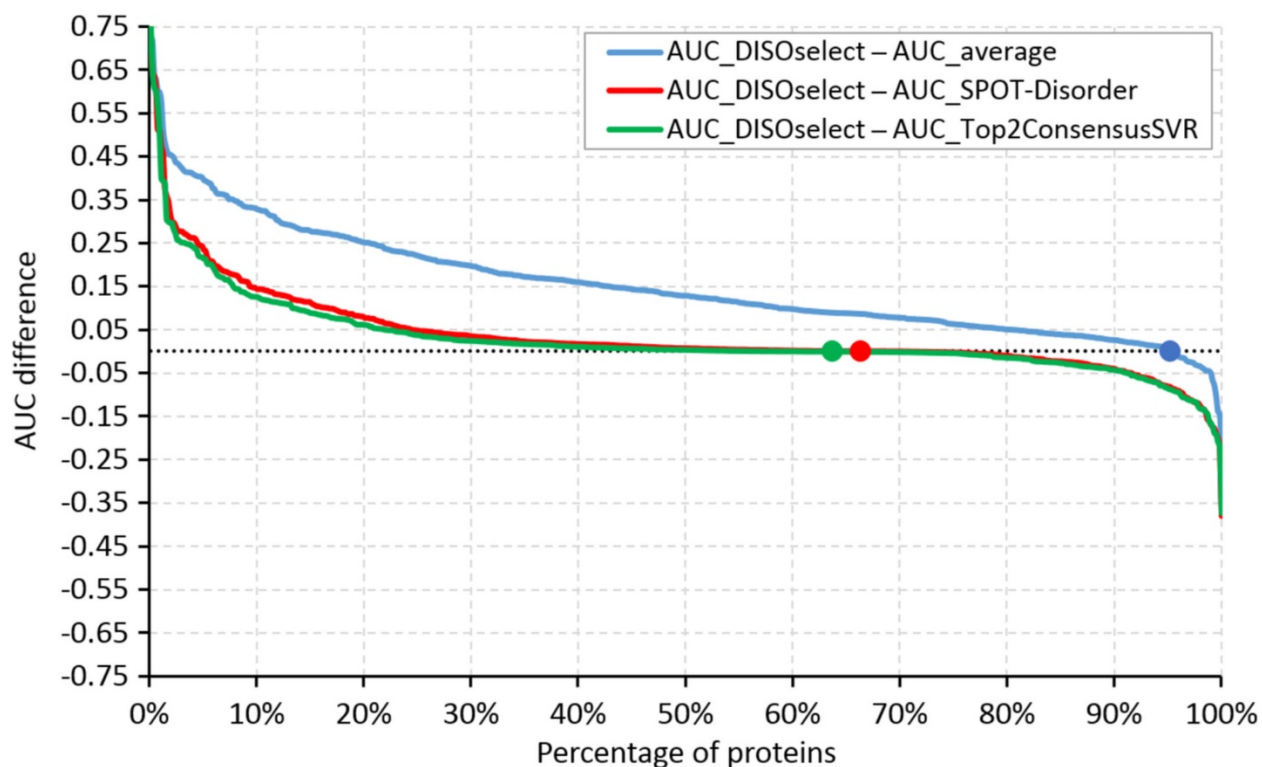


Figure 19: Evaluation of the differences in the protein-level area under the receiver operating characteristic curves (AUCs) for the same test proteins between the predictions selected with DISOselect and the average AUC of the 12 disorder predictors (blue line), between the predictions selected with DISOselect and the predictions generated by the most accurate disorder predictor at the dataset level, SPOT-Disorder (red line), and between the predictions selected with DISOselect and the best consensus-based method that relies on the support vector regression (SVR) (green line). Points indicate where the difference between protein AUCs crosses zero. The proteins are sorted by the value of the difference in the descending order. These results were published in [140].

4.4 DISOselect webserver

We developed and released a webserver that implements DISOselect as a free service for non-commercial users. This webserver is available at <http://biomine.cs.vcu.edu/servers/DISOselect/>. DISOselect requires only the FASTA-formatted protein sequences as input. Up to 1000 proteins can be predicted in a single run. All computations are performed on the webserver side. The webserver outputs the putative AUC and the qualitative performance (including the percentile of predicted AUC value) for each of the 12 disorder predictors, which are sorted in the descending order of the predicted AUC. The predictor at the top of the list, which has the highest estimated AUC, is recommended to the user as the best option to collect the disorder predictions. For the user's convenience, the main page of the webserver provides links to the websites of these

disorder predictors under the “Help” section. The results are available via an HTML page, which can be accessed via a direct link, and a parsable text file. We will archive these results for at least one month.

4.5 Summary

Our empirical analysis shows that the per-protein predictive quality of popular disorder predictors varies widely between different proteins. The users cannot expect that the disorder predictor with the best benchmark-dataset level results will provide favorable results across all proteins. These results suggest that a computational tool that can accurately estimate per-protein predictive performance for a given disorder predictor and a given protein is needed.

To this end, we developed a new recommendation system, DISOselect, which accurately identifies well-predicted proteins for each of the 12 considered disorder predictors and which recommends the best performing disorder predictor for a given input protein. We analyze these two capabilities from several different perspectives and we show that DISOselect outperforms the currently available solutions. More specifically, the disorder predictions selected using DISOselect are significantly more accurate than the results produced by any of the 12 disorder predictors, including the top-performing methods such as SPOT-Disorder and DISOPRED3, and a selection of four conventional consensus predictors. The average per-protein AUC for the predictions selected with DISOselect is 0.97, compared to an average AUC of 0.82 generated by the 12 methods, and an average AUC of 0.94 for the consensus methods.

DISOselect provides two key advantages to the end users. First, it offers advice on whether a given disorder predictor would provide an accurate prediction for a given protein of user’s interest. Second, if users are comfortable with using multiple disorder predictors, DISOselect accurately recommends the most suitable predictor. Importantly, besides suggesting the best tool, DISOselect informs the users about the expected predictive quality of this selected and other disorder predictors. This fast-to-compute insight is provided before the user has to actually make the possibly time-consuming disorder prediction.

Chapter 5. Assessment and comparative analysis of the predictive performance of disorder predictions for specific functional types of disordered proteins

Protein functions and structures can be classified using a wide range of resources, which include Pfam [157], SCOP [158], SUPERFAMILY [159], and CATCH [160]. Along those same lines, IDPs are classified based on different criteria, such as function [5, 25, 34, 161, 162], functional motifs [163], and sequence features [164, 165]. This chapter addresses objective 3, which aims to compare predictive performance of disorder predictions for specific functional types of IDPs. These results were published in [166].

As we discuss in Chapter 3, the predictive performance of disorder predictors was empirically assessed in numerous comparative studies [68-78]. The results of these studies, which were surveyed in Chapter 3 (at the protein level) and a recent article (at the dataset level) [14], can be used to perform side-by-side comparisons of the predictive performance of the disorder predictors. They can be used to guide the users to select accurate predictors and to inform the both users and developers about the current levels of predictive quality offered by the best tools. The latter fuels the progress in the development of gradually more accurate tools. This progress was recently summarized in [23], where the authors show that the predictive performance measured with AUC (Area Under the ROC Curve) have risen from the 0.73 to 0.79 range in mid 2000s, to the 0.85 to 0.90 range that is secured by the methods that were published in the last four years. While the past comparative studies provide invaluable insights, they also share a few drawbacks. First, they perform the assessment using generic sets of proteins while they rarely (only once) analyzed performance for specific functional protein families. Second, they overlook an important aspect of the similarity between the benchmark dataset and the training datasets that were used to develop the tested predictors. High levels of similarity may result in an overestimation of the predictive performance and may distort the results by favoring certain methods for which the similarity is higher. Third, some of the previous assessments assume that non-disordered regions from the proteins that are collected from DisProt database are structured

while some of these regions bear the possibility of being disordered. To this end, we performed the first-of-its-kind comparative analysis that includes analysis of the performance for several key functional types of disordered proteins (protein- and nucleic acid-binding proteins [167-170]), that develops and utilizes a new benchmark dataset that ensures controlled/low levels of similarity with the training datasets of the assessed predictors with validated structured regions.

5.1 Overview of the past intrinsic disorder predictor assessments

We start our assessment by providing a detailed overview of previous surveys of the intrinsic disorder predictors. We identified a total number of 28 surveys regarding intrinsic disorder predictors that were published during the period from 2003 to 2020 in peer reviewed venues [14, 68-78, 171-186]. The conventional format used in the above surveys is to start with a historical background of disorder predictor development and subsequently contrast the designs of different methods, including their input features and predictive models. In some of the surveys, selected disorder predictors were compared based on their predictive performance.

The first reported survey on disorder predictors was published in 2003 as part of the CASP5 [68] (Critical Assessment of Structure Prediction), which is 5th edition of the biannual assessment of the protein structure predictions. We summarize the abovementioned 28 surveys in the figure 20 based on their date of publication and categorize them into to three groups considering the intended target of the assessment. The respective categories include the assessments of disorder predictors, assessments of the disorder function predictors and assessments that consider both disorder predictors and disorder function predictors. The recent surveys put more emphasis towards assessing the disorder function predictors over the disorder predictors. Functions of IDRs are classified based on their cellular function and the interaction partners [187-189]. Recent study reveals that protein binding and nucleic acids binding collectively account for 84% of available molecular partner annotations in the DisProt for IDR [171]s. The specific fractions of available molecular partner annotations in the DisProt for IDRs with different interaction partners are as follows. Protein binding accounts for 66% of available molecular partner annotations while nucleic acid binding, metals, lipids, small molecules and inorganic salts respectively account for

17%, 6%, 5% ,5% and 1%. These fractions justify our focus on the protein binding and nucleic acid binding IDRs in this assessment.

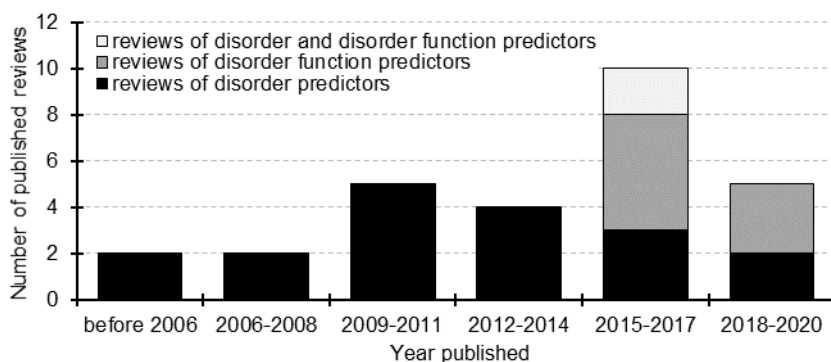


Figure 20: Chronological summary of the past surveys of the intrinsic disorder and intrinsic disorder function predictors. This figure was published in [166].

Out of the 28 surveys, 11 have conducted comparative assessments to compare the predictive performance between multiple disorder and disorder function predictors. We further analyze the above mentioned 11 surveys in Table 8 to assess their impact and scope. Majority of the above surveys accumulated over 100 citations according to Google Scholar, suggesting that they have attracted substantial amount of interest. These 11 surveys include six editions of CASP from CASP5 (2002) to CASP10 (2012) [68, 73, 75-78]. The disorder assessment category was discontinued from the CASP series after 2012 with the intention of initiating a separate dedicated community driven assessment for intrinsic disorder.

Table 8: Summary of the past comparative assessments of disorder predictors. The articles are sorted chronologically (from the most recent). The citation numbers were collected from Google Scholar on September 29, 2020. Predictors shown in the bold font in the “suggested best disorder predictors” column are included in the comparative assessment in this section. This table was published in [165].

Article	Target of assessment	Suggested best disorder predictors (year published)	Year assessment published	Year most recent assessed predictor published	Number of citations	Benchmark dataset has reduced similarity with training sets of the assessed predictors
this article	disordered proteins; disordered protein-binding protein; disordered nucleic acids-binding proteins	N/A	N/A	2018	N/A	yes
[69]	disordered proteins	SPOT-Disorder (2017), DISOPRED3 (2015)	2019	2017	4	no
[70]	disordered proteins	Espritz (2012)	2018	2017	33	no
[71]	disordered proteins	DisEMBL (2003), IUPred (2005)	2015	2012	121	no
[73]	disordered proteins	DISOPRED3 (2015), PrDOS (2007), MFDp (2010)	2014	2015	128	no
[72]	disordered integral membrane proteins	PreDisorder (2009)	2014	2012	12	no
[74]	disordered proteins	MFDp (2010), MD (2009), PONDR-FIT (2010)	2012	2010	149	no
[75]	disordered proteins	PrDOS (2007), DISOPRED (2004)	2011	2010	118	no
[76]	disordered proteins	GS-MetaServer (2012), PreDisorder (2009)	2009	2008	131	no
[77]	disordered proteins	DISOPRED (2004), DISpro (2005)	2007	2006	109	no
[78]	disordered proteins	predictor by Obradovic et al.	2005	2004	114	no
[68]	disordered proteins	N/A	2003	2002	97	no

Table 8 reveals that depending on the assessment, the set of the disorder predictors that are indicated to provide the best predictive performance differs. This inconsistency is partly due to the fact that these articles consider different sets of disorder predictors in the assessments. However, majority of the surveys suggested that the best disorder predictors are among the most recently published methods, respective to the time when the survey was conducted. As an example, the latest comparative survey that was published in 2019 [69] indicates that SPOT-Disorder [190] and DISOPRED3 [191] as best disorder predictors while SPOT-Disorder was commissioned after nine out of 11 surveys were published. The latest three surveys [69-71], at the time when this work was published in 2020 [166], indicate that SPOT-Disorder [190], DISOPRED3 [191], ESpritz [192], DisEMBL [54] and IUPred [193-195] as the highest performing disorder predictors.

According to Table 8, when it comes to the intended target of the assessment, 10 surveys assess the disorder predictions for a generic set of proteins while one [72] assesses the disorder predictions for the disordered integral membrane proteins. The latter survey compares the predictive performance among 13 disorder predictors using a dataset that consists of around 350 membrane proteins. Potential reason for the fact that none of the past survey focused on the disorder function predictors could be the limited number of available experimental annotation for the disorder functions. The disorder binding partner annotations were added to the DisProt database in 2016 and in limited quantities. They were further extended from the original list of 1108 IDRs to 1476 IDRs only in 2020 in the new release of DisProt.

We discuss another important aspect of disorder predictor assessments in the last column of Table 8 where we consider the sequence similarity of the test data they use with respect to the training datasets of the predictors that they asses. None of the 11 assessments take measures to limit the sequence similarity of the benchmark dataset against the training datasets. As an example, the CASP assessments [68, 73, 75-78] create their benchmark datasets using unreleased PDB structures at the time of the assessment and disregard the similarity to the training datasets of the considered predictors. Other recent surveys collect their benchmark proteins from MobiDB, DisProt and UniProt databases without limiting the sequence similarity to the training datasets of the assessed predictors [69-71]. Interestingly, the process of training and

validation/testing of new disorder predictors ensures that the sequence similarity of the test dataset against the training datasets is intentionally reduced, at least for the methods that is being proposed. The usual practice is to limit the sequence similarity between test dataset and training dataset below 30% [62, 196-200]. Limiting the sequence similarity of the test dataset is meant to demonstrate that a given method can solve the non-trivial problem of predicting dissimilar sequences. This is because a simple sequence alignment is capable of making strong predictions in the presence of high similarity.

Altogether, we show that past surveys share several drawbacks. They do not consider evaluating disorder predictions for specific disorder functions (except for one survey that focuses on membrane proteins) and fail to properly control sequence similarity of their test dataset to the training datasets of the tested methods. We address these aspects in our subsequently described survey.

5.2 Selection of disorder predictors

We cover a comprehensive set of 10 disorder predictors. The selection of these predictors was motivated by their availability, availability of their training datasets, computational efficiency and previously reported predictive performance. The selected computational disorder predictors cover all three types of prediction architecture classes that were described in the Section 2.2.1. The analysis includes GlobPlot [47], IUPred-short and IUPred-long [48] representing the *ab initio* methods. DisEMBL, VSL2B [53], SPOT-Disorder [100] that are the machine learning methods. Finally, DISOPRED3 [112] and the three versions of ESpritz that are tuned to predict intrinsic disorder annotated from X-ray structures (ESpritz-Xray), NMR structures (ESpritz-NMR) and using DisProt database (ESpritz-DisProt) [192] represents the meta predictors. These predictors are highly-cited and by extension often used. Their citation numbers are 1763 (IUPred), 1241 (DisEMBL), 1002 (GlobPlot), 680 (VSL2B), 355 (DISOPRED3), 256 (ESpritz), and 107 (SPOT-Disorder); source: Google Scholar on April 14, 2020. Moreover, this selection of methods overlaps with the predictors that were covered in the recent comparative assessments [69-71] and includes the five tools (SPOT-Disorder, DISOPRED3, DisEMBL, IUPred and ESpritz) that have been highlighted as the best-performing in the last three comparative surveys [69-71].

5.3 Collection of benchmark dataset

According to a recent analysis, the two most commonly annotated partners of IDPs are proteins and nucleic acids, which collectively cover 84% of the partner-annotated disordered regions in the DisProt resource [171]. More precisely, 66% partner-annotated regions have protein partners, 17% nucleic acids partners, 6% metals, 5% lipids, 5% small molecules, and 1% inorganic salt partners. This motivates our focus on the protein- and nucleic acids-binding IDPs.

Recent comprehensive survey concludes that an “updated and more comprehensive benchmark datasets should be established” [14]. Correspondingly, we establish a new benchmark set with two main objectives in mind: 1) explicit reduction in similarity to the training sets of the selected 10 disorder predictors; and 2) inclusion of the annotations of the protein-binding and nucleic-acid binding proteins. First, we collected the complete set of 1,418 proteins from the newest version 8 of DisProt that have experimental annotations of disorder and binding partners; we exclude the annotations marked as “ambiguous” in DisProt. Second, we collected the training datasets of the ten disorder predictors. We clustered the combined set of the DisProt and training proteins using CD-Hit [201] at 30% sequence similarity and we removed all clusters that include at least one training protein. The remaining set of 319 DisProt proteins is dissimilar to the training proteins (at 30%) and includes functional annotations that allow us to identify the sets of protein-binding proteins (that include at least one disordered protein-binding region) and nucleic-acid binding proteins (that include at least one disordered nucleic acid-binding region). Third, we mapped the un annotated regions from DisProt protein sequences into PDB. In order to execute this mapping, we first create a database of PDB sequences where regions that lack structure are masked. Then we align unannotated regions from Disprot protein sequences to above masked PDB sequence database using BLAST. We annotate any unannotated DisProt protein region which aligns to at least one masked PDB sequence region with $\geq 90\%$ similarity and have e-value ≤ 0.1 as structured. In order to do generate a more diverse and balanced dataset we match the number of fully disordered proteins (38) in the dataset with equal number of fully structured proteins from PDB. When we are selecting fully structured sequences from PDB we make sure to minimize the risk of them including any disordered regions by selecting monomers with high-quality crystal

structures (resolution < 2Å) that cover complete UniProt sequences based on mapping with SIFTS [202]. We cluster abovementioned fully structured protein sequences from PDB with the training datasets of selected disorder predictors using CD-HIT at 30% sequence similarity threshold and select 38 random proteins which do not belong to clusters with protein sequences from training datasets of selected disorder predictors. The final dataset consists of total number of 357 proteins that has experimental disorder and disorder function annotations from Disprot, validated structured region annotation from PDB and unmapped regions where they are excluded from the assessment. This benchmark dataset is summarized in Table 9. We emphasize that this dataset not only ensures low similarity to the training datasets of the tested methods but also uses high-quality annotations of the structured regions and includes fully structured proteins. This allows us to accurately study performance of the disorder predictors on the structured proteins and regions, which is another important feature that is missing in the past surveys.

Table 9: Summary of the benchmark dataset. This table was published in [165].

Metric	Complete dataset	Protein-binding proteins	Nucleic acids-binding proteins
Number of proteins	357	108	15
Number of residues	186,337	38,221	5,934
Number of disordered residues	31,608	14,125	1,567
Disorder content (% of disordered residues)	0.17	0.37	0.26

5.4 Comparative assessment of predictive performance

5.4.1 Effect of the sequence similarity reduction and structured region validation on benchmark dataset

In this section we assess the impact of the sequence similarity reduction and structured region validation of the benchmark dataset to the predictive performance of selected 10 disorder predictors. As an initial point of reference, we report the predictive performance of same 10 disorder predictors from previously conducted surveys. The previously reported predictive performance nine of the ten predictors was taken for from [70] and for the remaining SPOT-

Disorder from [203]. The previous evaluations do not limit the similarity of benchmark dataset against the training datasets of the respective disorder predictors as well as do not validate the experimental annotations for the unannotated regions in the DisProt proteins. In Figure 21, we denote the results from previous surveys as “previous results” using the black line. We also report results for two versions of our new benchmark dataset. One with limited similarity to the training datasets of the tested predictors (but without validation of the experimental annotations of order, like it was done in the past surveys) and in the other version we limit the similarity to the training datasets and validate the experimental annotations of unannotated regions in the DisProt proteins. The results for the methods in the first version are denoted as the “limited similarity benchmark” in Figure 21 and shown using the red line. The predictive performance for the second version of dataset with limited similarity and ordered region validation is denoted as the “new benchmark” and shown using green lines.

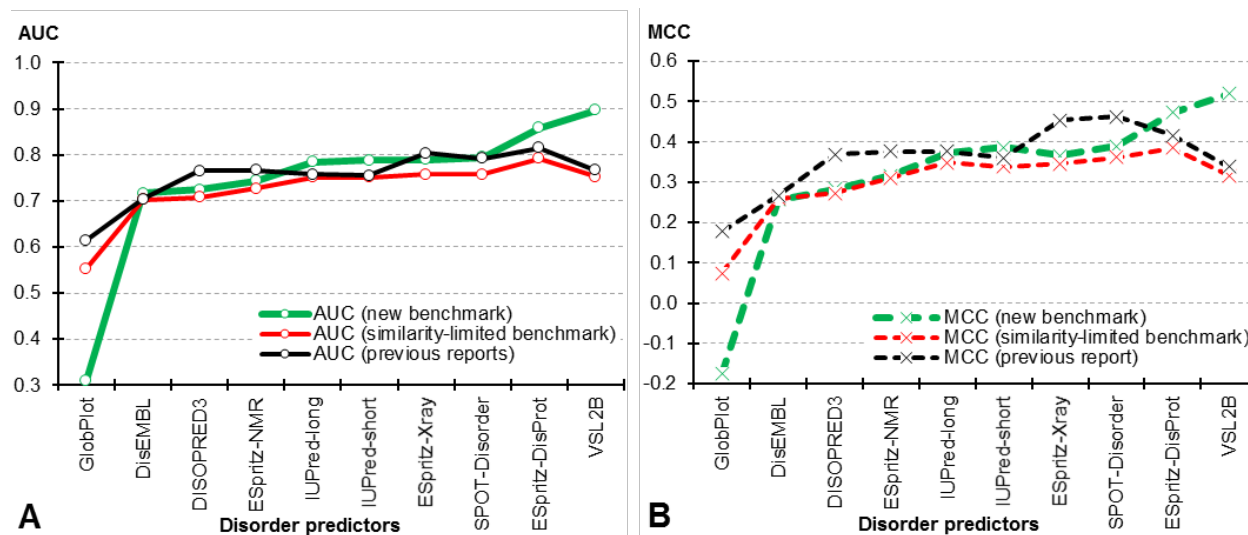


Figure 21: Comparison of the predictive quality measured with AUC (panel A; solid lines) and MCC (panel B; dashed lines). We report results on the new benchmark (in green; dataset with <30% sequence similarity to the training proteins + with experimental validation of structured regions + with fully structured proteins), based on recent previous reports (in black; datasets with no limits on sequence similarity to the training proteins + with no experimental validation of structured regions + with only disordered proteins), and based on a similarity-limited benchmark (in red; a version of the new benchmark dataset with <30% sequence similarity to the training proteins + no experimental validation of structured regions + only disordered proteins). The latter dataset is a proxy for the datasets used in prior studies with the only difference being the reduced similarity to the training proteins. Disorder predictors are sorted by their AUC values on the new benchmark dataset. This figure was published in [165].

The analysis of the differences in the predictive performance between the “previous results” (black lines in Figure 21) and the “limited similarity benchmark” (red lines in Figure 21) allows us to assess the effect of limiting sequence similarity against the training datasets of the selected disorder predictors. These results are highly correlated with Pearson’s correlation coefficient of 0.94 for ROC-AUC and 0.89 for MCC. Moreover, we note a consistent drop in the predictive performance across the 10 considered predictors from the “previous results” to the “limited similarity benchmark” results in both ROC-AUC and MCC. The average ROC-AUC across all 10 considered predictors drops by 0.03 (0.72 vs 0.75). The average MCC drops by 0.06 (0.03 vs 0.36). Furthermore, the best methods based on the two predictive performance measures also record a drop in the predictive performance with the sequence similarity reduction. ESpritz-DisProt which records the highest AUC in the past survey drops by 0.05 (from 0.804 to 0.758) and SPOT-Disorder which records the highest MCC in the past survey drops by 0.10 (from 0.462 to 0.361). These observations demonstrate that results on the benchmark datasets that share high sequence similarity with the training datasets of the predictors that they assess tend to consistently inflate the predictive performance.

The predictive performance difference of methods between the “limited similarity benchmark” and the “new benchmark” experiments demonstrate the impact of providing experimental validation to the unannotated regions in the DisProt proteins. We observe that the predictive performance improves when using the validated experimental annotations of the ordered regions. According to Figure 21, the highest improvement is for VSL2B which improves ROC-AUC by 0.14 and MCC by 0.20. In contrast, the overall poorly performing GlobPlot drops its predictive performance when using the experimentally validated order annotations. GlobPlot is originally designed to differentiate globular proteins from non-globular proteins, where the lack of “globularity” acts as a proxy to detect disorder. We show that GlobPlot detects validated structured regions as non-globular while they are in fact not disordered (and not necessarily globular).

Table 10: Predictive performance on the new benchmark dataset. The table lists results on the complete benchmark dataset with 357 proteins, the set of 38 fully disordered proteins, the set of 38 fully structured proteins, and the benchmark dataset of 319 proteins that exclude the fully structured proteins. We quantify statistical significance of differences in AUC between the best predictor (identified in bold font) and each the other nine predictors on a given dataset. We bootstrap 50% of the proteins 100 times. For normal measurements (tested with the Anderson-Darling test at 0.05 significance) we use the paired t-test; otherwise we use the Wilcoxon rank sum test; = and + mean that the differences are not significant (p-value > 0.01) and significant (p-value ≤ 0.01), respectively. This table was published in [165].

Predictor	Benchmark dataset					Fully disordered proteins	Fully ordered proteins	Benchmark dataset without fully ordered proteins				
	AUC	Precision	Sensitivity	FPR	MCC	Sensitivity	FPR	AUC	Precision	Sensitivity	FPR	MCC
VSL2B	0.897	0.609	0.845	0.204	0.519	0.925	0.000	0.805+	0.611	0.845	0.399	0.404
ESpritz-DisProt	0.858+	0.593	0.487	0.060	0.473	0.811	0.052	0.842	0.685	0.487	0.067	0.486
SPOT-Disorder	0.795+	0.334	0.756	0.261	0.390	0.662	0.290	0.826+	0.578	0.756	0.234	0.485
ESpritz-Xray	0.790+	0.375	0.623	0.193	0.366	0.702	0.226	0.812+	0.586	0.623	0.160	0.459
IUPred-short	0.788+	0.431	0.613	0.170	0.386	0.692	0.176	0.801+	0.607	0.613	0.165	0.444
IUPred-long	0.785+	0.422	0.693	0.233	0.373	0.834	0.262	0.806+	0.625	0.693	0.206	0.463
ESpritz-NMR	0.743+	0.336	0.721	0.310	0.317	0.774	0.351	0.776+	0.563	0.721	0.272	0.414
DISOPRED3	0.724+	0.294	0.653	0.293	0.283	0.662	0.340	0.767+	0.513	0.653	0.248	0.380
DisEMBL	0.717+	0.308	0.439	0.162	0.257	0.559	0.193	0.741+	0.520	0.439	0.132	0.336
GlobPlot	0.310+	0.122	0.428	0.655	-0.175	0.388	1.000	0.563+	0.332	0.428	0.326	0.096

5.4.2 Comparative assessment of disorder predictors on the benchmark dataset

In Table 10 we compare the predictive performance of the ten disorder predictors. The left panel gives the results for the full benchmark dataset with 357 proteins. VSL2B predictor records the highest ROC-AUC of 0.89 which it is significantly better than other 9 predictors (p -value <0.01). At the same time several other disorder predictors record high predictive performance (AUC >0.75 and MCC >0.35) on this benchmark dataset including ESpritz-DisProt, SPOT-Disorder, ESpritz-Xray and both versions of IUPred. Precision measures of the VSL2B and ESpritz-DisProt reveal that these methods identify majority of the disordered residues correctly. Moreover, VSL2B and SPOT-Disorder record the top two sensitivity values indicating they identify over 75% of the native disorder residues. These results reveal that several disorder predictors are capable of providing accurate predictions.

The right panel of Table 10 reports the predictive performance of the ten selected methods on a subset of the complete benchmark dataset where we exclude fully structured proteins. In this subset ESpritz-DisProt shows the highest predictive performance with ROC-AUC of 0.85 which is significantly better than other nine methods (p -value <0.01). The sensitivity remains same as the complete benchmark dataset as the native disordered residues are unchanged. The precision improves by a large margin for all predictors, except for VLS2B, when the fully ordered proteins are removed. This is because majority of these methods predict substantial number of false positives in the fully ordered proteins. This is further validated by the FPR column for the fully ordered proteins where the considered methods predict between 5% (ESpritz-DisProt) and 100% (GlobPlot) of the false positive predictions. The one exception is VSL2B that produces no false positives in these proteins. At the same time, we observe that FPRs for the fully ordered proteins are in line with FPRs in the subset of the dataset that includes disordered proteins. We find that nine out of ten methods overpredict disorder in the fully ordered proteins as well as in the ordered regions of the disordered proteins. The one exception, VSL2B, scores 0.00 FPR in the fully ordered proteins and highest sensitivity in the fully disordered proteins. This explains VLS2B's

overall best predictive performance in our similarity-limited benchmark dataset (Table 10 and Figure 21).

We conclude that VSL2B is the most versatile method among the ten considered disorder predictors, particularly when it comes to predicting disorder for the fully ordered and fully disordered proteins. Apart from this, ESpritz-DisProt and SPOT-Disorder provide strong predictive quality measured with the ROC-AUC and MCC scores.

5.4.3 Predictive performance assessment on the disordered protein-binding and nucleic acid-binding proteins

In this section we provide the comparative analysis of predictive performance for the 10 disorder predictors in two main functional sub-classes of the disordered proteins, namely the disordered protein-binding proteins and the disordered nucleic acid binding proteins. We compare the predictive performance of the methods across the above mentioned two functional sub-classes as well as against the complete benchmark dataset. In order to conduct this comparison, we equalize the native disorder content among the two sub-classes and the complete benchmark dataset. The disorder content equalization is motivated by several previous studies that show that the predictive performance of disorder predictors drops when tested on proteins with larger amount of the native disorder [17, 59, 69, 126].

We perform disorder content equalization by subsampling the two larger sets (the complete benchmark dataset and the disordered protein binding subset) to match the disorder content in the smallest set of the disordered nucleic acid binding proteins. We start the subsampling by calculating the protein level native disorder content distributions for each dataset and quantifying the significance of the differences between these distributions. Next, we remove the proteins from two large subsets that increase the p -value by largest margin until we reach p -value of 0.001. Figure 22 shows that predictive performance distribution of the 10 methods before (grey box plots) and after (white box plots) we equalize the native disorder content. Figure 22 shows that even though absolute values of the predictive performances are shifted by disorder content equalization, the relative differences in the predictive performance across the

datasets remain similar (i.e., differences between the white box plots follow the same pattern as the differences for the gray box plots).

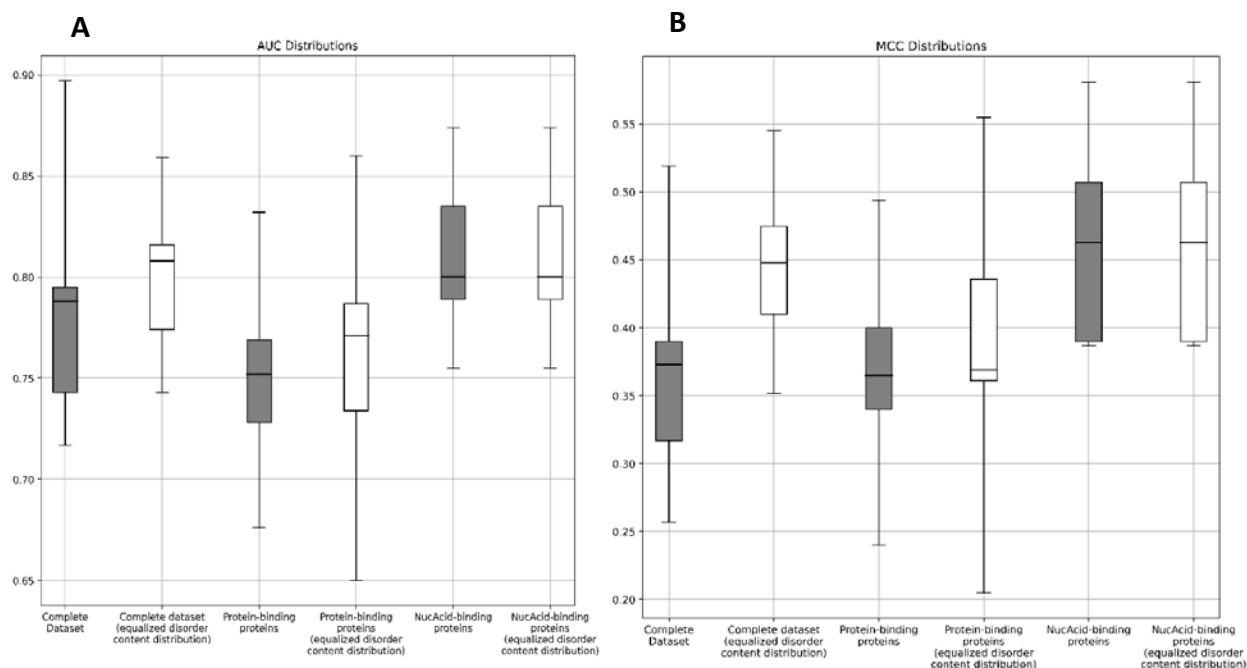


Figure 22: Distribution of the AUCs (panel A) and MCCs (panel B) over nine disorder predictors. We exclude the poorly-performing GlobPlot from this analysis. The box plots show the lowest AUC (bottom error bar), first quartile (bottom of box), median (horizontal line inside box), third quartile (top of box) and highest AUC (top error bar). The grey plots are for the original datasets while the white plots are for the sampled/disorder content-equalized datasets that have similar distribution of the per-protein disorder content. The content distribution similarity was measured using Kolmogorov–Smirnov test at p-value of 0.001. This figure was published in [165].

Figure 23 shows the predictive performance of the ten disorder predictors across two functional sub-classes of disorder and for the generic disordered proteins in the benchmark dataset. These results reveal that predictions for the disordered nucleic acid binding proteins secure similar levels of predictive performance as the performance for generic disordered proteins. The average (across predictors) ROC-AUC is 0.781 for the nucleic acid binding proteins while it is 0.774 for the generic disordered proteins. Similarly, the corresponding average MCCs are 0.422 and 0.406, respectively. At the same time, we observe that quality of the predictions for the disordered protein binding proteins drops substantially when compared to other two results, with the average ROC-AUC of 0.739 and average MCC of 0.356. According to Figure 23, this decrease in

the predictive performance is consistent across eight of the ten predictors. One of the exceptions is ESpritz-DisProt that maintains similar levels of predictive performance across two functional sub-classes and the generic disordered proteins. The second exception is GlobPlot that registers consistently poor predictive performance. The best performing disorder predictor for the generic disordered proteins and disordered protein binding proteins is ESpritz-DisProt while for the nucleic acid binding proteins the best results are produced by SPOT-Disorder. These results are in agreement with a recent study that was conducted using structured/ordered proteins [204]. It shows that the quality of the predictions of the protein binding residues is much worse than the quality of the predictions of the DNA/RNA binding protein residues for the structured proteins.

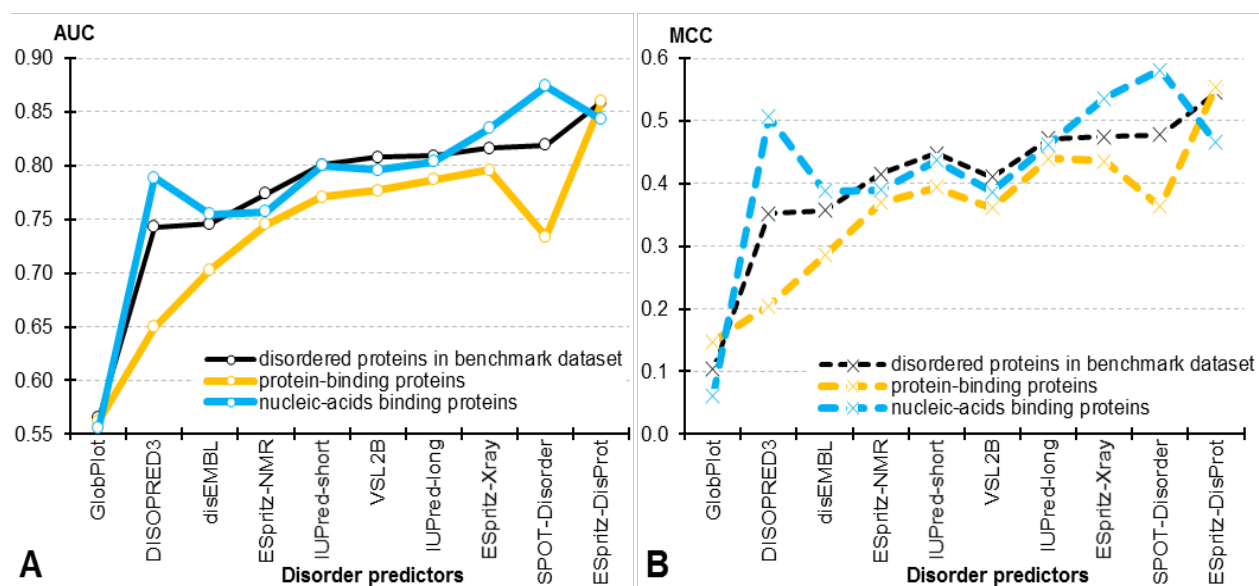


Figure 23: Comparison of the predictive quality measured with AUC (panel A; solid lines) and MCC (panel B; dashed lines). We report results on the generic set of disordered proteins (i.e., proteins that have disordered residues) from benchmark dataset (in black), the disordered protein-binding proteins (in yellow), and the disordered nucleic acids-binding proteins (in blue). Disorder predictors are sorted by their AUC values on the disordered proteins. This figure was published in [165].

5.5 Summary

Prediction of intrinsic disorder plays an important role to bridge the large and growing gap between the limited amount of the available experimental annotations and the large numbers of newly discovered unannotated protein sequences. Over 60 disorder predictors are currently available. This large number of tools makes it challenging to identify the best disorder predictor

for a given set of proteins. One option to solve this problem is to use DISOselect, the tool that we developed and describe in Chapter 4. Another option is to rely on the disorder predictor assessments that focus on providing insights that would facilitate selection for specific classes of the disordered proteins. This would be suitable in scenarios where the end user would know the class of the disordered proteins that they work with (e.g., whether it bind proteins or nucleic acids). We identify 11 surveys that provide comparative assessment of predictive performance [68-78] and observe that only one of them provides analysis for a specific class of disordered proteins (membrane proteins). To this end, we conduct first-of-its-kind analysis that assesses the predictive performance for two large functional sub-classes of the disordered proteins: protein binding and nucleic acid binding proteins. Our analysis also overcomes two other limitations of the previous studies: use of test datasets that may include proteins that are similar to the training datasets of the tested predictors and lack of validation for the annotations of the ordered regions. We consider a representative set of ten disorder predictors and a benchmark set with limited similarity to the training datasets of these predictors and we use validated experimental annotations of the ordered regions in that test dataset.

Our assessment reveals that limiting the similarity of the benchmark dataset against the training datasets of the selected disorder predictors results in a consistent and substantial drop in the predictive performance when compared with previous assessments that did not limit the similarity [69-71, 73, 205]. This means that the results of the past studies likely overestimate the predictive performance.

The proper validation of the experimental annotations of the ordered regions in the disordered proteins results in higher levels of the reported predictive quality. This is because higher quality ground truth information is used to quantify the predictive performance. We suggest that future assessments should take this into consideration when designing benchmark dataset.

We identify three disorder predictors that offer particularly strong predictive performance. VSL2B is the most versatile method that provides the best results for the fully structured and fully disordered proteins and very strong results for proteins that have disordered regions. However, ESpritz-DisProt and SPOT-Disorder outperform VSL2B for the latter type of proteins. Most

importantly, we analyze performance for the two functional classes of disordered proteins: protein and nucleic acid binding proteins. We demonstrate that majority of the disorder predictors offer much lower predictive performance for the disordered proteins with protein binding regions. The AUC and MCC values for these proteins are on average lower by 0.04 and 0.05 when compared to the set of generic disordered proteins, and by 0.04 and 0.07 when compared to the disordered nucleic acids-binding proteins, respectively.

These results suggest certain functional classes of disordered proteins are more difficult to predict accurately, calling for the development of new disorder predictors. At the same time, our research also shows complete lack of tools to predicts some of the functional classes of the disordered proteins. While specialized tools for the predictions of the disordered protein and nucleic acids binding are already available [172, 173, 206], we find that disordered lipid binding interactions are devoid of predictive tools. This predictive target was neglected primarily because of the limited availability of reliable experimental annotations. However, with the recently growing numbers of experimental annotations and rapid advancements in machine learning techniques that allow building models from limited-size datasets (e.g., deep neural networks coupled with transfer learning), the development of computational methods to predict these neglected functions become feasible. The next chapter addresses the development of such new predictive tool.

Chapter 6. Accurate prediction of the disordered lipid-binding residues from protein sequences

6.1 Introduction and motivation

This chapter addresses objective 4 that is defined in Section 1.2. We have conceptualized, developed and comparatively tested a runtime-efficient computational system that accurately predicts disordered lipid-binding residues in the intrinsically disordered proteins. IDRs are shown to have molecular level interactions with multiple binding partners including proteins, DNA, RNA, lipids, metal ions and other small molecules [34, 167, 168, 207-211]. However, only few hundred of IDRS are experimentally annotated with their binding partners [34, 206]. These data can be used to build tools capable of predicting IDR-partner interactions for the majority of the disordered proteins that currently lack these functional annotations [172, 173, 180, 206, 212].

As we show in Section 2.2.2 (Table 4), significant majority of current computational predictors of disorder functions aim to predict disordered protein-binding regions. In contrast, there is only one method, DisoRDPbind [104, 105] for the prediction of the DNA and RNA interactions. To best of our knowledge, there are no tools that predict disordered lipid-binding regions. The lack of the computational methods that predict interactions with majority of the binding partners can be explained by the lack of a sufficient amount of the experimentally annotated data to develop and validate such predictors. However, recent rapid growth in the amount of the available experimental annotations for the binding partners has enabled the development of new computational predictors. As an example, the latest version (version 8.0) of the DisProt, which is the largest repository of the functional annotations of IDPs, has grown by about 50% in the amount of the annotations for the disordered lipid binding regions when compared to the previous version of DisProt (version 7.2) [8, 33].

Lipid molecules carry out many important structural and functional roles including energy storage, regulation, signaling, insulating and transporting [213-218]. Studies show that lipids interact with proteins. For instance, lipid molecules facilitate fibrillogenesis by inducing protein

structures to assemble into protofibrillar and fibrillary structures [219-222]. Conventional experimental techniques to identify protein-lipid interactions include immunocytochemistry, cytotoxicity assays, circular dichroism spectroscopy, calcein leakage and differential scanning calorimetry [223-226]. Several diseases associated with misfolding of IDPs have a connection with their affinity to bind lipid molecules [227]. Misfolding of α -synuclein, which is a fully disordered protein, and tau proteins, which includes significant amount of disorder, are examples where disorder and protein-lipid interactions are connected with pathogenic conditions like the Parkinson's disease, Alzheimer's disease, multiple-system atrophy, and dementia with Lewy body [224, 225, 228-231]. As another example, SecA from *E. coli* illustrates an interaction between IDRs and a lipid bi-layer [232]. Moreover, some bacteriocins, such as colicin A, unfold to the disordered molten globule state when they interact with the cytoplasmic lipids of the host cell to perform membrane insertion [233].

Motivated by the recent growth in the annotations of the lipid-interacting IDRs and the functional importance of these interactions, we present DisoLipPred, first-of-its-kind predictor of the disordered lipid binding residues (DLBRs). These residues are intrinsically disordered, interact with lipids and exclude the transmembrane regions. This means that DisoLipPred produces predictions that complement the results generated with the current predictors of the transmembrane regions [234-236]. DisoLipPred utilizes a deep neural network to predict propensity for lipid binding in disordered regions for each amino acid in the input protein sequence. The design of this tool relies on several innovations. First, we utilize transfer learning. We start with a more generic network that predicts IDRs that interact with different types of partner molecules, which is motivated by the large amount of the underlying training data. We freeze this partner type-agnostic network and extend it to develop the final model that specializes the predictions to the lipid partners. Second, we use literature to identify physiochemical properties that are associated with protein-lipid interactions and use them to expand the inputs to the deep network. Third, we deploy a new training and prediction strategy that bypasses ordered/structured residues. More specifically, we train the deep network models using only the native disordered residues to identify DLBRs. This focuses our model on identifying DLBRs among other disordered residues, compared to a more traditional scenario

that differentiates DLBRs from both structured and disordered residues. During the prediction process we use a modern disorder predictor to identify disordered residues which are processed by our deep network to predict DLBRs. The predicted ordered residues bypass the network, since by default they exclude DLBRs. We perform ablation analysis that empirically demonstrates that these innovations lead to significant improvements in the predictive performance when compared to a more traditional design that exclude these solutions. Such traditional design is characteristic to the current predictors of the IDRs that interact with proteins and nucleic acids [92, 102, 105, 195, 206, 237-240].

6.2 Materials and methods

6.2.1 Dataset description

We collect experimental data to establish training and test datasets. We use the training dataset to design and train the deep network. More specifically, we further subdivide the training dataset into learning and validation subsets where we use the learning partition (2/3 of the training dataset) to train the model which we test on the validation partition (1/3 of the training dataset). We exclude the test set from the training process and use it solely to perform comparative assessment against other, indirect approaches to predict DLBRs. These datasets are composed of three types of proteins: proteins with DLBRs, proteins with other IDRs and fully structured proteins. This allows us to develop and test models that differentiate DLBRs from other disordered and structured residues. We collect the proteins with IDRs and DLBRs from version 8 of DisProt [8]. We exclude disordered regions with an ambiguous function or structure annotations, which are tagged in DisProt. We also exclude the proteins with IDRs that do not have annotated function to minimize the likelihood of false negative annotations (some of these IDRs could bind lipids). Moreover, inspired by recent works [166, 205], we further process the proteins from DisProt to ensure that we use high-quality annotations of structured regions. Instead of assuming that regions that lack disorder annotations are by default structured, we map the un-annotated regions to the sequences of the protein structures from Protein Data Bank (PDB) [241], for which we mask the disordered residues. We utilize the

protocol from [166] that relies on the alignment with Basic Local Alignment Search Tool (BLAST) algorithm [155]. The regions in the DisProt sequences that share >90% similarity and e-value <0.1 with at least one masked PDB sequences are assumed structured. We collect the fully structured proteins from PDB [241]. We minimize the likelihood that these proteins include IDRs by collecting high-resolution (<2Å) monomers that do not have disordered regions (i.e., structure is resolved for all amino acids) and which map into full UniProt sequences based on SIFTS [202]. We cluster the combined collection of these three protein types using the CD-HIT algorithm with 25% similarity [242]. We place the entire clusters into either training or test datasets, which ensures that these datasets share <25% sequence similarity. The test dataset is composed of half of the proteins with DLBRs, 100 proteins with other IDRs and 100 structured proteins. We place the remaining proteins into the training dataset. We use two versions of the training dataset to implement the transfer learning. The complete training, which we use to generate the partner type-agnostic deep network, includes the proteins with DLBRs, proteins with IDRs that interact with other molecules and structured proteins. We transfer this network into an expanded network that predicts DLBRs and which we train using a resampled training dataset, which we dub target training dataset. This dataset focuses on the proteins with DLBRs by undersampling at random 100 proteins with the other IDRs and 100 structured proteins. We provide details of these datasets, including their overall sizes and numbers of annotated residues, in the Table 11.

Table 11: Description of the training and test datasets.

Dataset	Number of residues			Number of proteins	
	disordered lipid binding	disordered	all	fully structured	all
Complete training dataset	1,921	141,018	2,426,416	1,446	2,892
Target training dataset	1,921	17,823	96,015	100	211
Test dataset	1,471	20,623	106,348	100	219

We use a secondary test dataset to empirically assess whether DisoLipPred’s predictions of DLBRs in fact exclude the transmembrane regions. We sourced this TM (transmembrane) test dataset from a recent study that introduced SCAMPI2 predictor of the transmembrane regions [234]. We clustered the transmembrane proteins used in that study together with the proteins

from the complete training dataset using CD-HIT at 25% similarity and selected the transmembrane proteins from clusters that exclude the training proteins. We combine these transmembrane proteins with the transmembrane proteins from the test dataset to devise the TM dataset. This dataset includes 25 proteins, 15,978 amino acids and 4,308 transmembrane spanning residues, and shares <25% sequence similarity to the training datasets.

6.2.2 DisoLipPred architecture

The prediction workflow of the DisoLipPred consists with four main components as shown in Figure 24, namely bypass module, sequence profile module, neural networks and rescaling module. The input protein sequence is first processed by SPOT-Disorder [198], one of the most accurate disorder predictors according to multiple recent assessments including the CAID experiment [69, 166, 205]. The SPOT-Disorder's predictions are fed into the bypass module that separates the predicted disordered residues, which are subsequently processed by the deep network to predict DLBRs, from the predicted order residues, which bypass the deep network prediction. Next, sequences of proteins with the predicted disordered residues are used to derive sequence profiles. The profiles incorporate sequence-derived structural and functional information that is relevant to the prediction of DLBRs. They are utilized as the input to a deep neural network that predicts propensity for disordered lipid binding and which is designed using transfer learning. Finally, the rescaling module normalizes and merges the outputs from the deep network with the predictions of the ordered residues from the bypass module, producing the final predictions.

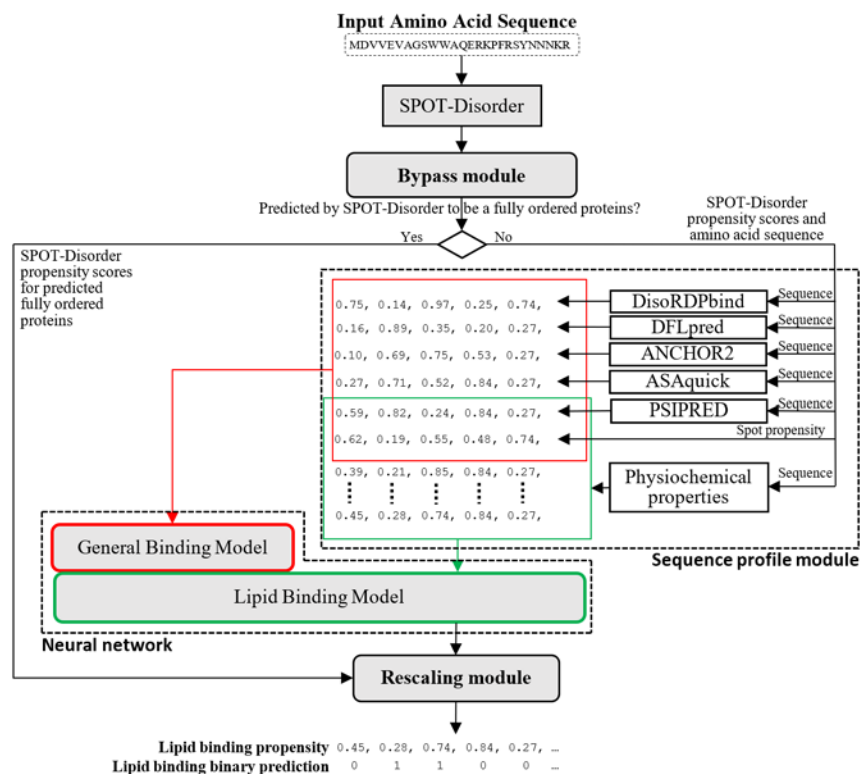


Figure 24: Prediction workflow of DisoLipPred.

6.2.2.1 Bypass module

DLBRs are localized in the disordered regions. The main challenge for DisoLipPred is to identify these lipid-binding residues among the other disordered residues. Consequently, during the training process we train and validate the deep network on the native disordered residues. We exclude the ordered residues from training since they can be accurately identified with one of the currently available accurate disorder predictors. We use the highly-accurate SPOT-Disorder predictor [198] for that purpose. The bypass module separates disordered from ordered residues based on the SPOT-Disorder's predictions, such that the putative ordered residues bypass the prediction process while the putative disordered residues are selected for prediction with the deep network. The SPOT-Disorder generated propensities for the putative ordered residues are rescaled and combined with the deep network generated propensities in the rescaling module to produce the propensities for DLBRs. We use ablation analysis to demonstrate that the approach that applies the bypass module provides more accurate results than the direct prediction of DLBRs from all residues.

6.2.2.2 Sequence profiles

The sequence profiles provide a rich source of information that is relevant to the prediction of DLBRs and derived directly from the sequences. We use two profiles to facilitate the transfer learning. One for the partner-agnostic portion of the deep network that aims to predict interacting disordered residues (red areas in Figure 24) and the other for the part of the deep network that predicts DLBRs (green areas in Figure 24).

The partner-agnostic profile relies on a comprehensive collection of predictors of structure, intrinsic disorder and disorder functions, with particular focus on the prediction of the interacting disordered regions. We use the predictions of the solvent accessibility from ASAquick [243], secondary structure from PSIPRED [147], disorder from SPOT-Disorder [198], protein, DNA and RNA interacting disordered regions from DisoRDPbind [104, 105], protein-binding disordered regions from ANCHOR 2 [195], and disordered linker regions from DFLpred [107]. This profile is summarized in Appendix 2.

The second profile, which serves as the input to predict DLBRs, focuses on the sequence-derived information that is specific to the lipid-binding. We use two relevant structural properties, the putative solvent accessibility and secondary structure generated with ASAquick [243] and PSIPRED [147], respectively, putative disorder from SPOT-Disorder [198], and a curated set of 46 physiochemical properties of amino acids that are associated with protein-lipid interactions [244]. These properties were selected empirically from a comprehensive collection of over 530 physiochemical indices from the AAindex database [245] based on their ability to discriminate between lipid-binding and non-lipid binding proteins [244]. They include hydrophobicity, hydrophobic moment, charge, isoelectric point, transfer energy, Gibbs energy, solvation free energy, propensity for helical and sheet conformations, and propensity for side chain interactions. Complete list of these properties is in Appendix 2

6.2.2.3 Transfer learning of the deep recurrent neural network model

Transfer learning is a training strategy where knowledge learned from a source domain/dataset is transferred to a related target domain/dataset to improve the learning in the target domain

[246]. This strategy is deployed when the target dataset has limited amount of data compared to a more data-rich source dataset, and is particularly useful for training the data-hungry deep neural networks [247]. Transfer learning was recently applied to predict secondary structures of RNA [248], caspase and metalloprotease cleavage sites [249], MHC-I peptide binding [250] and transcription factor binding [251], but it was never used to develop predictors of interacting disordered regions. Prediction of DLBRs offers an ideal scenario for the transfer learning. While we have a relatively limited amount of DLBRs (3,392 residues), the amount of the data concerning a generic set of interacting IDRs is very substantial (161,641 residues). Thus, we first build a partner-agnostic deep network using the complete training dataset, which we then freeze and extend with additional layers to develop the target network that predicts DLBRs using the target training dataset. We adopt deep recurrent networks given their recent success with the prediction of disorder [198, 199, 203].

The partner-agnostic network consists of two long term-short memory layers that are sandwiched between fully connected layers with ReLu activation function in the internal layers and the sigmoid activation function at the output layer (Figure 25A). We use the RMSprop optimizer, binary cross entropy as the loss function, dropout rate of 0.5 (to minimize overfitting), and dynamic adjustment of the learning rate which we set to gradually decrease as the training progresses. This network uses the partner-agnostic profile as the input. We optimized the number of layers and the number of neurons per layer using an iterative approach where we start from a small size and increase it by a small increment until AUC measured for the prediction of interacting IDRs on the validation set decreases in two consecutive iterations.

The optimized partner-agnostic network is transferred to develop the target network. We remove the output layer from the partner-agnostic model and freeze it. We connect the last layer of this network to several additional layers that narrow down the partner-agnostic prediction to the partner-specific prediction of DLBRs. This network extends the partner-agnostic profile with the additional inputs relevant to the prediction of DLBRs that we discuss in Section 6.2.2.2. This extension includes multiple bidirectional long short-term memory layers placed between fully connected layers (Figure 25B). Similar to the training of the partner-

agnostic network, we optimize the size of the additional layers using the increment approach that maximizes AUC for the prediction of DLBRs on the validation set.

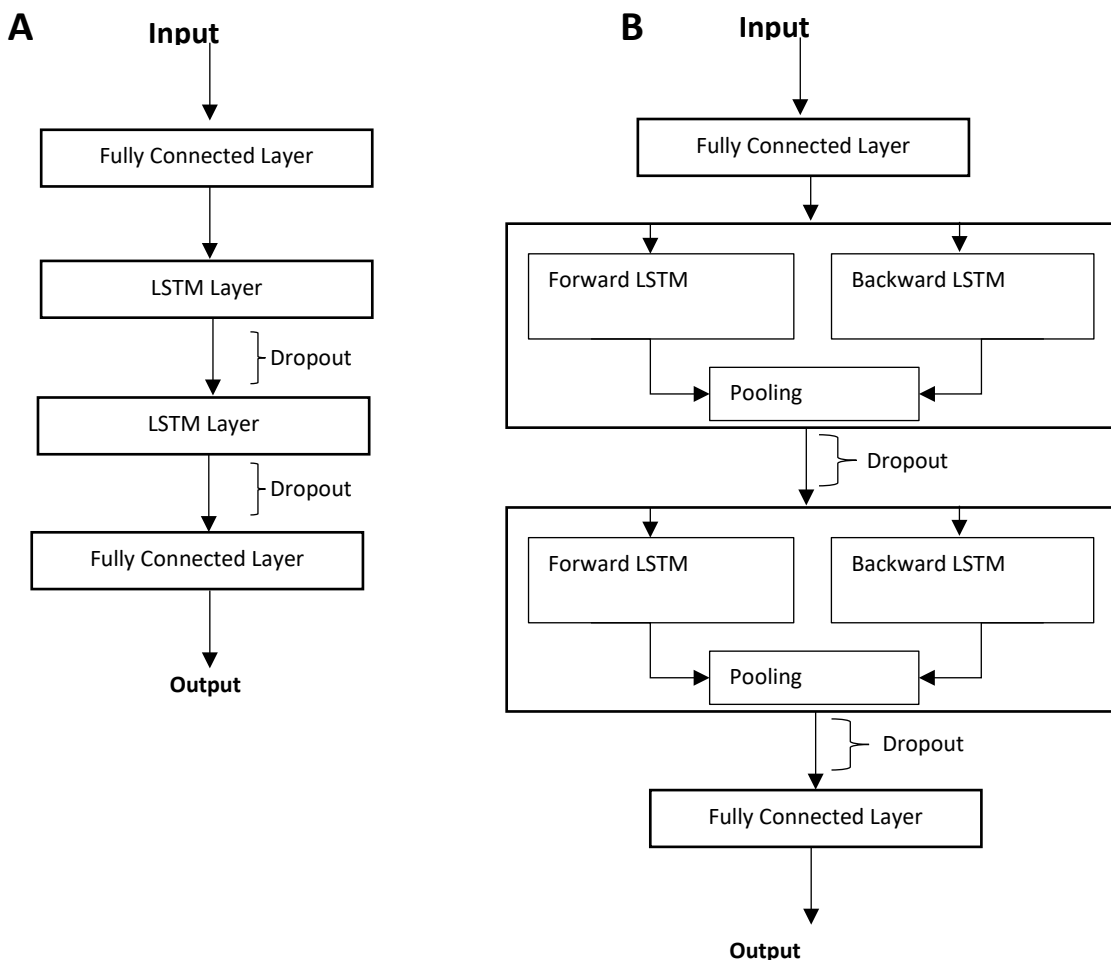


Figure 25: Architecture of the deep recurrent neural network used by DisoLipPred. Panel A shows the partner-agnostic network that we train using the dataset of IDRs that interact with different partner types. Panel B gives the network that extends the partner-agnostic network to perform the partner-specific prediction of DLBRs.

6.2.2.4 Rescaling module

We combine the disordered lipid-binding propensities generated by our deep recurrent neural network for the disordered residues predicted with SPOT-Disorder and the SPOT-Disorder's propensity scores for the predicted ordered residues. First, we normalize the outputs from the deep neural network to the unit range. We also rescale the SPOT-Disorder's propensities for predicted ordered residues, which bypass the neural network, so they cover the 0 to 0.5 range.

This aims to minimize risk of missing out DLBRs among the incorrect predictions of order from SPOT-Disorder. This way, these false negatives can be predicted with moderately high scores.

6.3 Results

6.3.1 Ablation analysis of the prediction model

The three main innovations underlying DisoLipPred include the use of the transfer learning, lipid-binding features and the bypass module. We perform ablation analysis to quantify the impact of these innovations on the predictive performance of DisoLipPred. To do that, we compare the results produced by the DisoLipPred model with the three setups where one of these features is removed and the setup where all three features are removed (Table 12). For instance, in the setup 1 we exclude transfer learning by removing the partner-agnostic network and relying solely of the lipid binding neural network. The bypass module works by training and testing the deep network on the disordered residues and sidestepping the deep network predictions for the putative ordered residues. The training process utilizes the native disordered residues while during tests/predictions we use the predictions from SPOT-Disorder. In setup 3, we evaluate the impact of using the predicted disordered residues for both training and testing/predictions. The setup 4 excludes all three innovations where for the bypass feature we train/test the deep network using both disordered and ordered residues. This bare-bone predictor is comparable to current deep learners that are used to predict disorder [196, 198, 203] and the protein binding IDRs [252, 253]. We trained each of the five setups separately by maximizing the AUC on the validation set.

Table 12: Experimental setups for the ablation study.

Setup	Use of transfer learning	Use of lipid features	Bypass module during training
DisoLipPred	Yes	Yes	Native disorder vs native order
1	No	Yes	Native disorder vs native order
2	Yes	No	Native disorder vs native order
3	Yes	Yes	Predicted disorder vs predicted order
4	No	No	No

Table 13: Predictive performance of DisoLipPred and its variants from the ablation analysis (Table 1) on the test dataset. We perform the assessment on the complete test dataset, and also on the subset of disordered residues from the test dataset. We quantify the binary metrics (sensitivity and F1) at the fixed specificity = 0.9. We assess the statistical significance of the differences between the results produced by DisoLipPred and each of the variants using procedure explained in Section 2.2. * indicates that DisoLipPred provides significantly better result (p -value < 0.05).

Setup	Complete test dataset				Disordered residues in the test dataset			
	AUC	Sensitivity	Specificity	F1	AUC	Sensitivity	Specificity	F1
DisoLipPred	0.781	0.382	0.900	0.145	0.635	0.286	0.900	0.201
1	0.747*	0.290*	0.900	0.112*	0.572*	0.162*	0.900	0.118*
2	0.745*	0.327*	0.900	0.125*	0.603*	0.146*	0.900	0.175*
3	0.726*	0.260*	0.900	0.101*	0.593*	0.177*	0.900	0.129*
4	0.678*	0.123*	0.900	0.049*	0.396*	0.046*	0.900	0.035*

We compare predictive performance of the five setups on the test dataset in Table 13. We assess the predictions on the complete datasets as well as on the subset of the disordered residues. The latter evaluation quantifies the ability of these models to solve a more difficult problem of identifying DLBRs among other disordered regions i.e., DLBR are more similar to other disordered residues than to the ordered residues.

DisoLipPred offers accurate predictions with AUC = 0.78 and sensitivity = 0.38. This sensitivity is relatively high given that we measure it at the low FPR = 0.10 (specificity = 0.90). Compared to the complete DisoLipPred model, we note a noticeable and statistically significant drop in the predictive performance for all metrics and ablation variants (p -value < 0.05). Among the setups where one of the innovations is removed, the largest drop is for the setup 3 where we manipulate the bypass feature. This suggests that our deep networks can be better trained to recognize DLBR among native disordered residues than among the predicted disordered residues. The errors from the disorder predictions and the networks training seem to

accumulate in the latter case. The results further substantially decline when all three innovations are removed (setup 4). This means that the contributions of the novel design features are complementary.

As expected, tests on the native disordered residues (right side of Table 13) lead to lower predictive performance across all methods. However, DisoLipPred still provides reasonably accurate predictions (AUC = 0.64 and sensitivity = 0.29 at FPR = 0.1). The ablation variants consistently underperform compared to the complete model (p -value < 0.05), with the bare-bone model performing at the random levels: AUC < 0.55 and sensitivity and F1 near 0. This demonstrates that the basic deep network is incapable of predicting DLBRs since it can only solve the trivial problem of differentiating DLBRs from ordered residues (AUC = 0.68 on the complete dataset vs. 0.40 on the disordered residues). In other words, the three innovations that we introduce are essential to provide accurate predictions.

6.3.2 Comparative assessment on the test dataset

We compare DisoLipPred to current alternatives that can be indirectly used to predict DLBR. We consider three categories of the indirect predictors. First, we include methods that predict transmembrane regions in protein sequences. We select predictors with publicly available implementations/servers that include one recently released method, SCAMPI 2 [234], and one older and highly-cited method, Phobius [235]. While DLBRs predicted by DisoLipPred exclude transmembrane regions, we investigate whether the transmembrane region predictors could be used to also predict DLBRs. Second, we cover disorder predictors since DLBR are one of the functional subtypes of the disordered residues. We choose 10 disorder predictors that were considered in recent comparative surveys [69, 166]: DisEMBL-465 (trained using X-ray structures) and DisEMBL-HL (trained to predict disorder-like loop conformations) [54]; three versions of ESpritz [65]: ESpritz-Xray (trained on X-ray structures), ESpritz-NMR (trained on NMR structures) and ESpritz-DisProt (trained on the DisProt database data); two flavors of IUPred [52, 193]: IUPred-short (trained to predict short IDRs) and IUPred-long (trained to predict long IDRs); GlobPlot [54] and SPOT-Disorder [198]. Third, we include representative predictors of disorder function, such as DisoRDPbind [104, 105] that predicts the disordered

RNA binding, DNA binding and protein binding residues, ANCHOR 2 [195] that predicts disordered protein binding residues and DFLpred [107] which predicts disordered linkers. Finally, we compute a baseline results based on sequence alignment to the training proteins. We perform this alignment with BLAST [155], where DLBR annotations are transferred from the aligned positions in the most similar training proteins that secures e-value < 1.0. We setup the e-value parameter to maximize performance on the test dataset.

Table 14: Predictive performance on the test dataset. We perform the assessment on the complete test dataset, and also on the subset of the native disordered residues from the test dataset. We quantify the binary metrics (sensitivity and F1) at the fixed specificity = 0.9. for the predictors that produce the propensity scores. We use the default sensitivity, F1, and specificity values for the other three methods that produce only binary predictions: SCAMPI 2, Phobius and BLAST. We assess the statistical significance of the differences between the results produced by DisoLipPred and every other tool using procedure explained in Section 2.2. * indicates that DisoLipPred provides significantly better result (p-value <0.05). Methods are sorted in the ascending order by their AUC within each predictor type group.

Predictor Type	Name	Complete test dataset				Disordered residues in the test dataset			
		AUC	Sensitivity	F1	Specificity	AUC	Sensitivity	F1	Specificity
Transmembrane regions	SCAMPI 2	N/A	0.019*	0.016*	0.98	N/A	0.019*	0.035*	0.99
	Phobius	N/A	0.016*	0.024*	1.00	N/A	0.016*	0.031*	1.00
Baseline	BLAST alignment	N/A	0.000*	0.000*	1.00	N/A	0.000*	0.000*	1.00
Disorder function predictors	DFLpred	0.338*	0.037*	0.015*	0.90	0.554*	0.109*	0.081*	0.90
	DisoRDPbind-RNA	0.450*	0.035*	0.014*	0.90	0.517*	0.028*	0.022*	0.90
	ANCHOR	0.637*	0.229*	0.090*	0.90	0.446*	0.178*	0.129*	0.90
	DisoRDPbind-Protein	0.556*	0.016*	0.006*	0.90	0.276*	0.002*	0.001*	0.90
	DisoRDPbind-DNA	0.636*	0.211*	0.083*	0.90	0.554*	0.062*	0.047*	0.90
Disorder predictors	GlobPlot	0.530*	0.225*	0.088*	0.90	0.482*	0.167*	0.123*	0.90
	ESpritz-NMR	0.571*	0.216*	0.085*	0.90	0.412*	0.113*	0.084*	0.90
	disEMBL-465	0.610*	0.119*	0.048*	0.90	0.433*	0.048*	0.037*	0.90
	disEMBL-HL	0.619*	0.143*	0.056*	0.90	0.477*	0.066*	0.050*	0.90
	IUPred-long	0.626*	0.256*	0.100*	0.90	0.420*	0.167*	0.123*	0.90
	IUPred-short	0.632*	0.257*	0.100*	0.90	0.441*	0.142*	0.105*	0.90
	ESpritz-Xray	0.659*	0.114*	0.046*	0.90	0.428*	0.070*	0.053*	0.90
	VSL2B	0.673*	0.205*	0.081*	0.90	0.433*	0.057*	0.045*	0.90
	SPOT-Disorder	0.692*	0.155*	0.062*	0.90	0.361*	0.043*	0.033*	0.90
	ESpritz-DisProt	0.768*	0.355*	0.135*	0.90	0.498*	0.065*	0.049*	0.90
DLBR predictor	DisoLipPred	0.781	0.382	0.145	0.90	0.635	0.286	0.201	0.90

Table 14 compares DisoLipPred’s predictive performance against the indirect predictors and the baseline. We derive the binary predictions from the propensity scores using thresholds that we

adjust to set FPR = 0.1 (specificity = 0.9). This allows us to directly compare the other binary metrics (sensitivity and F1) between methods. DisoLipPred provides accurate predictions of DLBRs on the test dataset, with AUC = 0.78 and sensitivity = 0.38 at FPR = 0.10. The latter means that DisoLipPred offers 3.8-fold increase in the rate of correct to incorrect predictions. Tests of statistical significance of differences reveal that the DisoLipPred's predictions are significantly better than the results of all 17 indirect methods and the baseline (p -value < 0.05). The poor performance of the baseline alignment stems from the low sequence similarity, < 25%, between the training and test proteins. The most accurate of the indirect predictors include Espritz-DisProt (AUC = 0.77, sensitivity = 0.35), SPOT-Disorder (AUC = 0.69, sensitivity = 0.16), and VSL2B (AUC = 0.67, sensitivity = 0.21). The ROC curves the test dataset for the best-performing methods, including DisoLipPred, SPOT-Disorder, VSL2B, Espritz-DisProt, are available in the Supplementary Figure S in Appendix 21 They reveal a large margin of improvement for DisoLipPred, particularly for low values of FPRs, i.e., conservative predictions where rate of false positives is low. We highlight the results from the two predictors of transmembrane regions that secure near zero (0.02) sensitivity at 0.1 specificity, which means that they do not predict DLBRs. We conduct a further comparison with the two transmembrane region predictors in 6.3.4 using a separate dataset of proteins with transmembrane regions.

The relatively high AUCs of several disorder predictors on the test dataset can be explained by the fact that they accurately differentiate DLBRs from the ordered residues. However, the results computed on native disordered residues in the test dataset (the right side of Table 14) reveal that these methods cannot reliably discriminate DLBRs from the other disordered residues. More specifically, AUCs of the top disorder predictors, Espritz-DisProt, SPOT-Disorder, VSL2B, are 0.50, 0.36 and 0.43, respectively. Overall, only DisoLipPred generates accurate results on the disordered residues while the other predictors are significantly worse (p -value < 0.05) and their performance is near random levels (AUC < 0.55). This is expected since none of the indirect tools were designed to predict DLBRs.

6.3.3 DisoLipPred predictions on the *Saccharomyces cerevisiae* proteome

We apply DisoLipPred to predict DLBRs for the complete *Saccharomyces cerevisiae* proteome that we source from UniProt [254]. The Baker's yeast proteome includes 6,049 protein sequences and 2,936,363 residues. This is one of the best-annotated proteomes; BUSCO (Benchmarking Universal Single-Copy Orthologs) scores its completeness at 99.6% [255]. We calibrate the binary predictions to 0.48% prediction rate (% putative DLBRs in the genome), which corresponds to the rate of the native DLBRs in the DisProt database. We exclude the putative DLBRs if they form segments of < 6 consecutive residues since the shortest experimentally annotated disordered lipid binding regions in DisProt are 6 residues long. We share these predictions on the DisoLipPred's website at <http://biomine.cs.vcu.edu/servers/DisoLipPred/>. We predict that about 4.9% of the yeast proteins have putative DLBRs (Figure 26A). Majority of these proteins have less than 5% of residues predicted as DLBRs, however, about 0.7% of the yeast proteins have a substantial amount of over 5% DLBRs (Figure 26B).

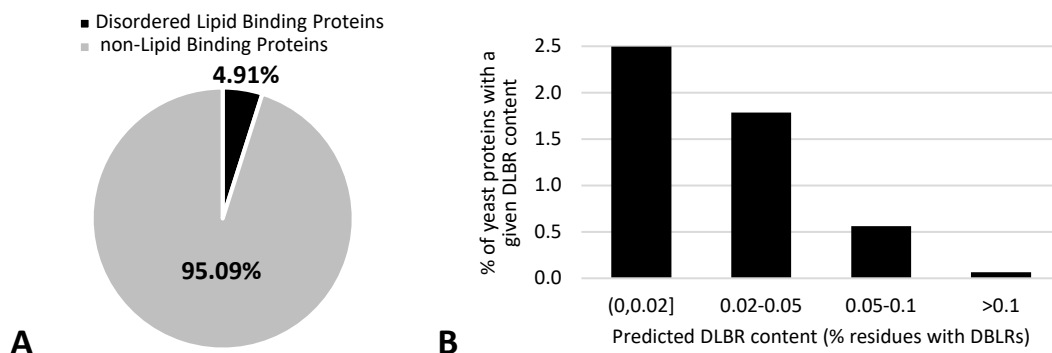


Figure 26: Summary of the DisoLipPred's predictions on the *Saccharomyces cerevisiae* proteome. Panel A shows the fraction of the yeast proteins predicted to have DLBRs. Panel B is the histogram of the putative content of DLBRs for the 4.9% of the yeast proteins with DLBRs.

We validate these predictions using the gene ontology (GO) annotations from UniProt. These annotations are independent of the ground truth data used in the test dataset. First, we select a subset of the yeast proteins that include the "lipid" keyword in their molecular function GO term and the "membrane" keyword within their cellular component GO term. The resulting set of 309 proteins is likely to be enriched in the proteins that have DLBRs; we call it GO lipid

associated protein set. Second, we compute the rate of proteins predicted to have DLBRs in the GO lipid associated protein set using DisoLipPred and compare it to the rate of these predictions generated with the second-best ESpritz-DisProt method (Table 14). We calibrate the ESpritz-DisProt’s predictions the same way as the predictions from DisoLipPred. Third, we calculate the expected rate of proteins with the putative DLBRs in the yeast proteome. We compute the rate for a randomly selected set of 309 yeast proteins and repeat this experiment 100 times to establish distribution of the expected rates. The results are summarized in Figure 27.

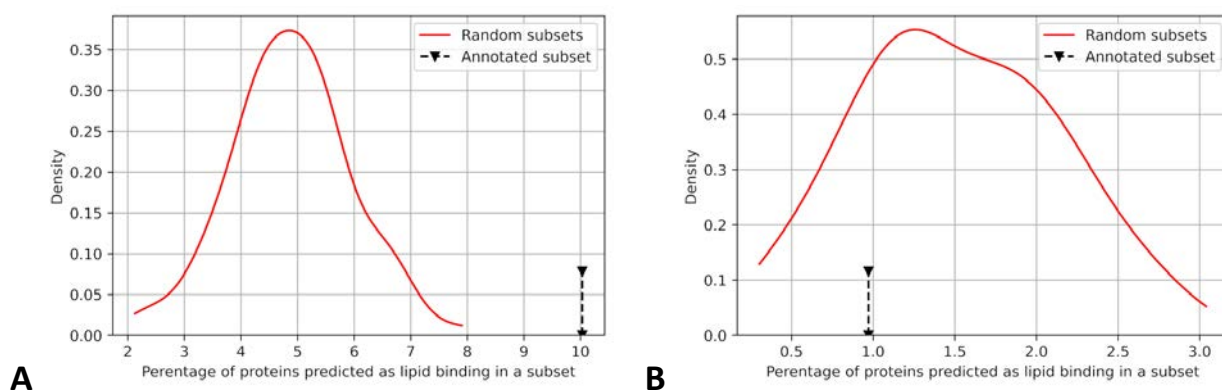


Figure 27: Analysis of the DisoLipPred predictions (Panel A) and the ESpritz-DisProt predictions (panel B) for the yeast proteins. The black arrows identify the rate of the putative proteins with DLBRs in the GO lipid associated protein set (i.e., set of 309 yeast proteins that share “lipid” keyword in the molecular function GO term and the “membrane” keyword in the cellular component GO term). Red lines show the distributions of the expected rates of the putative proteins with DLBRs, which we establish based on measuring the rate for 100 randomly selected sets of 309 yeast proteins.

The mean of the distribution for DisoLipPred’s predictions is 4.9% (Figure 27A) and corresponds to the overall rate of proteins with DLBRs in yeast (Figure 26A). DisoLipPred predicts 10.3% of proteins in the GO lipid associated protein set as having DLBRs. This rate doubles the expected rate of 4.9% and the difference is statistically significant based on the distribution of the expected values in Figure 27A (p -value < 0.01). On the other hand, the calibrated predictions from ESpritz-DisProt identify only 0.97% of the GO lipid associated protein set as having DLBRs. This rate is below the expected rate of the ESpritz-DisProt’s predictions (red line in Figure 27B), for which median is 1.5%. This suggests that the GO lipid associated proteins are overall depleted in disorder. In spite of the disorder depletion, the rate of the DisoLipPred’s predictions

of DLBRs is $10.3/0.97 = 10.6$ times higher than the rate of the ESpritz-DisProt's predictions, providing further support for our claim that DisoLipPred's predictions are accurate.

6.3.4 DisoLipPred prediction assessment on transmembrane proteins

Given that DLBR are defined as disordered lipid-binding regions that exclude transmembrane segments, we empirically evaluate whether the DisoLipPred's predictions in fact exclude the transmembrane residues. We test DisoLipPred and the two representative predictors of the transmembrane regions, SCAMPI 2 [234] and Phobius [235], on the TM dataset (Table 15). We introduce the TM dataset in the section 6.2.1. Here, we use the predictions from these three tools to identify native transmembrane regions, i.e., transmembrane residues are set as the positives while the other residues, including a small amount of DLBRs, are set as negatives. Since SCAMPI 2 and Phobius produce only binary predictions and thus their prediction rate cannot be calibrated, we adjust the rate of the DisoLipPred's predictions to match the specificity of each of the two transmembrane predictors. Table 15 shows that as expected SCAMPI 2 and Phobius provide accurate predictions of the transmembrane regions based on their high sensitivity scores, i.e., 0.79 sensitivity at the low 0.09 FPR and 0.57 sensitivity at the low 0.06 FPR, respectively. Their predictive positive rate (PPR) defined as the rate of true positives among the predicted positives is also relatively high and equals 0.28 and 0.20, respectively. In stark contrast, DisoLipPred's sensitivity values calibrated to the rate of predictions from SCAMPI 2 and Phobius are 0.04 and 0.03, demonstrating that it predicts very few transmembrane residues as DLBRs. These values are substantially smaller than the corresponding sensitivity values on the test dataset (Supplementary Figure S1 in Appendix 2). DisoLipPred's PPR is higher than its corresponding sensitivity because several proteins in this dataset include DLBRs, which by definition do not overlap with transmembrane regions. Altogether, these results show that DisoLipPred accurately differentiates between the transmembrane regions and DLBRs. Moreover, given the correspondingly low sensitivity of SCAMPI 2 and Phobius for the prediction of DLBRs (Table 14), we conclude that DisoLipPred predicts lipid interacting residues that complement the results produced by the predictors of the transmembrane regions.

Table 15: Predictive performance on the TM dataset. The performance is measured assuming that the native transmembrane regions constitute positive annotations. Both transmembrane predictors (SCAMPI 2 and Phobius) produce only binary predictions and thus their prediction rate cannot be calibrated. Instead, we calibrate the rate of the DisoLipPred's predictions to match the specificity of SCAMPI 2 and Phobius.

Predictor	Sensitivity	Specificity	PPR	F1
SCAMPI 2	0.795	0.91	0.279	0.780
DisoLipPred at SCAMPI 2 specificity	0.041	0.91	0.076	0.063
Phobius	0.574	0.94	0.197	0.663
DisoLipPred at Phobius specificity	0.035	0.94	0.050	0.059

6.3.5 Case study

We illustrate the DisoLipPred's predictions for one of the test proteins, the Sec-independent protein translocase protein TatA (UniProt accession number: P69428). Our objective here is to visualize and explain the predictions, rather than to evaluate their performance. TatA is a membrane associated protein, which is a subunit of the larger twin-arginine translocation (Tat) system [226]. The Tat system acts as a facilitator to transport large folded proteins through cellular membranes by creating a protein conducting channel [256, 257]. TatA contains a long IDR (positions 21 to 89) which was characterized with NMR [258]. Furthermore, proton based NMR revealed that part of this IDR (positions 21 to 44) binds to lipids [226]. Figure 28 shows DisoLipPred's predictions for TatA along with the abovementioned native annotations of the disordered and disorder lipid binding regions. DisoLipPred generates relatively high propensities at the N terminus half of the protein, resulting in the prediction of a long segment of DLBRs that overlaps with the experimentally determined lipid-binding region. Interestingly, we predict that the residues at the N terminus are also lipid binding. DisProt does not offer a conclusive evidence whether this segment is disordered or structured. Our alignment-based mapping into PDB (see Section 6.2.1) did not identify a known structure for this segment. Further investigation of literature reveals support for our prediction, where this segment is shown to likely interact with lipids of the cell membrane from the cytoplasmic side [259]. Altogether, this prediction agrees with the experimental annotations and provides support for

the hypothesis that the disordered lipid binding region is larger than DisProt suggests, covering the N-terminus half of the TatA sequence.

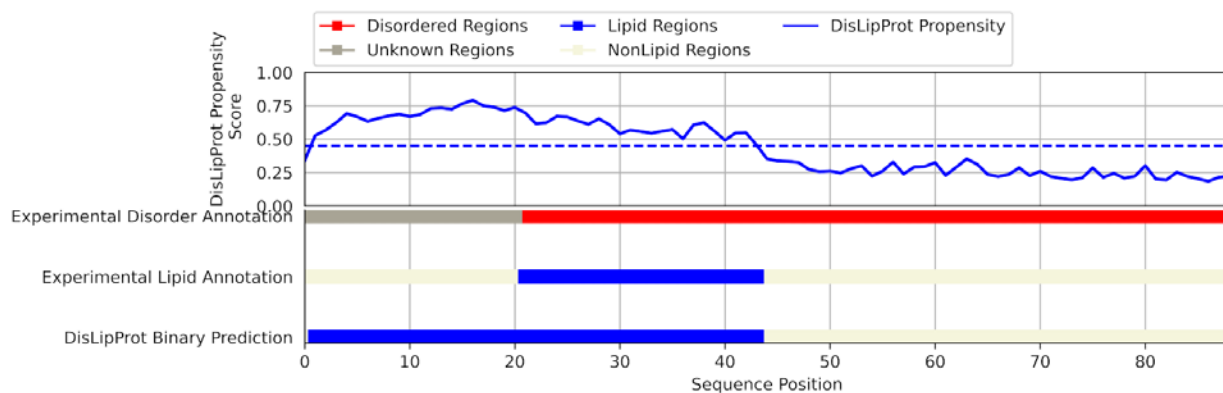


Figure 28: DisoLipPred predictions for TatA protein (Uniprot: P69428; DisProt: DP00834). The blue line in the top panel shows the residue level propensity scores generated by DisoLipPred. The horizontal blue bars at the bottom are the corresponding experimental annotation of lipid binding regions and the binary prediction from DisoLipPred. The horizontal red bar shows the experimental annotation of the intrinsic disorder, where grey color identifies regions that lack disorder/order annotations.

6.3.6 Webserver

DisoLipPred is freely available as a webserver at <http://biomine.cs.vcu.edu/servers/DisoLipPred/>. DisoLipPred webserver takes up to two amino acid sequences in the FASTA format as the input. The entire prediction process is automated, done on the server side and takes about 2-4 minutes for an average size protein. Users can optionally provide an email address where we send a notification email with the of the unique URL of results once the prediction is completed. The webserver provides the output propensities and binary predictions for each amino acid in the input protein sequence(s). The threshold that we use to generate the binary predictions corresponds to the 10% FPR on the training dataset. The outputs are available in two complementary formats: as a parsable text file and an interactive figure. The figure provides a graphical summary of the predictions with the zoom in/ out functions, ability to hide user-selected panels and take a screenshot, and mouse hover that shows additional information.

6.3.7 Summary

IDRs interact with many partner molecules including proteins, RNA, DNA and lipids. Sequence-based prediction of these IDRs is currently possible for the interactions with proteins and nucleic acids [171-173, 180]. Motivated by the growing amount of experimental data and the need to expand coverage of the current predictors, we have designed, implemented, validated and deployed a novel computational approach, DisoLipPred. This first-of-its-kind predictor accurately identifies DLBRs within intrinsically disordered protein regions. Our solution implements three innovative features that include the application of transfer learning, bypass module and selected physiochemical properties associated with protein-lipid interactions.

We deliver a multifaceted validation of the predictions produced by DisoLipPred. The ablation tests show that the quality of the DisoLipPred predictions is fueled primarily by the three innovations. Analysis on the test dataset reveals that DisoLipPred generates accurate predictions and that current tools that could be indirectly used to identify DLBR cannot differentiate the lipid-interacting residues from the other disordered residues. Validation on the complete yeast proteome provides further support for the claim that DisoLipPred produces accurate results. Moreover, we demonstrate empirically that the DisoLipPred's predictions complement the results produced by the predictors of the transmembrane regions. Altogether, our analysis suggests that DisoLipPred provides high-quality predictions of DLBRs that complement the currently available tools. DisoLipPred is available via a convenient webserver at <http://biomine.cs.vcu.edu/servers/DisoLipPred/>.

Chapter 7. Summary

This thesis focuses on multiple aspects related to the computational prediction of the intrinsic disorder and disorder functions. We started with an innovative study that evaluates the predictive performance of disorder predictors at their protein level. This analysis suggests that the predictive performance of disorder predictors should not be adjudicated solely based at the dataset level results. This is because the protein-level results are very different from what the dataset-level assessments suggest. Our analysis that spans a dozen disorder predictors shows that the protein level predictive performance has a consistent wide distribution with long tails at the lower range of predictive performance. This shape of distribution reveals that majority of protein level predictions should have higher than expected quality (higher than the protein-level results). However, a large number of proteins at the long tails are predicted very poorly. We also reveal that the predictive performance of disorder predictors is correlated with the amount of the native disorder content. To be more specific, the performance of disorder predictors substantially drops for proteins with higher disorder content. Finally, our study suggests that disorder predictors provide complementary results. This means that none of them is capable of providing the “silver-bullet” solution.

The objective 2 investigates the possibility of building a recommender system that would suggest a suitable disorder predictor for a given protein sequence. This objective is motivated by the diversity of the protein-level predictive performance values within and across current disorder predictors, which we identified under objective 1. We present a novel framework that predicts expected predictive performance of a given disorder predictor for the input protein, and uses these results collected over a set of disorder predictors to select the most accurate method. We use the physiochemical and putative structural properties of the input protein sequence to generate accurate prediction of the predictive performance. Extensive empirical tests demonstrate that our recommender system, DISOselect, significantly outperforms all current solutions, which include 12 representative disorder predictors and selected classical meta-models. We provide DISOselect to the research community as a webserver.

The third objective investigates the predictive performance of a selected set of 10 representative disorder predictors for two key functional subclasses of IDPs: disordered protein binding proteins and disordered nucleic acid binding proteins. We introduce three novel aspects in this analysis to address drawbacks of the past disorder predictor evaluations. We are the first to create a benchmark dataset with test proteins for which similarity was reduced with respect to the training datasets of predictors that we evaluate. Second, we use experimentally validated annotations of the ordered regions. Finally, we compare the predictive performance for a generic set of disordered proteins with the performance on the disordered protein binding and the disordered nucleic acid binding proteins. Our analysis reveals that similarity reduction of the benchmark dataset with respect to the training datasets results in a substantial reduction of the predictive performance when compared to the performance reported in previous studies where benchmark datasets do not limit the similarity. This suggests that the prior studies overestimate the predictive equality of disorder predictors. We also show that use of the experimentally validated ordered regions has a positive impact on the measured predictive performance. This stems from the fact that the ground truth concerning the order annotations has higher quality. Most importantly, we reveal that disorder predictions for the disordered nucleic acid binding proteins are accurate and share similar quality with the predictions for generic disordered proteins. However, disorder predictions for the disordered protein binding proteins suffer lower predictive performance, suggesting that future disorder predictors should focus on optimizing predictions for these proteins.

In the fourth and final objective we design and develop DisoLipPred, the first method that predicts disordered lipid-binding residues (DLBRs) in protein sequences. We introduce three innovations in the development of DisoLipPred. First, we use transfer learning based deep recurrent neural network. The transfer learning utilizes related large data concerning ligand-agnostic interactions in IDRs to facilitate/kickstart the design of the network module responsible for the predictions of DLBRs. We also utilize the bypass training strategy to train the model using native disordered regions and use it to make predictions for the predicted disordered residues. These predictions are combined with the predictions of the ordered residues. The main purpose behind the bypassing strategy is to improve the neural network's ability to identify DLBRs among

native disordered residues. The third innovation identifies physiochemical properties of proteins which are associated with lipid binding using literature and uses them to enhance inputs to the deep neural network. Ablation analysis shows that these three innovations are the main drivers behind the ability of DisoLipPred to accurately predict DLBRs. We also evaluate the DisoLipPred's predictive performance against multiple alternative/indirect prediction methods on an independent test dataset (i.e., test proteins are dissimilar to the proteins in the training dataset). This test shows that DisoLipPred is the only tool that provides accurate predictions of DLBRs. Moreover, we show that DisoLipPred does not cross-predict transmembrane protein regions as DLBRs using a separate test dataset of the transmembrane proteins. Comparison of DisoLipPred against two representative predictors of the transmembrane regions shows that these tools provide complementary results. We apply DisoLipPred to make predictions on the complete yeast proteome. This prediction facilitates additional evaluation using the gene ontology annotations, which further confirms that DisoLipPred provides accurate results. Lastly, we provide DisoLipPred to the research community as a publicly available webserver.

The work covered in the first three objectives was recently published in reputed peer-reviewed journals [125, 140, 166].

7.1 Major contributions

My major contributions related to specific objectives are as follows.

Objective 1: Elucidation and comparative analysis of protein-level predictive performance for current disorder predictors.

- Evaluated the predictive performance of disorder predictors at individual protein level and contrasted it to their dataset level performance.
- Investigated the complementarity of predictive performance across selected set of 13 computational disorder predictors both at the protein and dataset levels.
- Investigated the impact of disorder content of individual proteins to their predictive from selected set of 13 computational disorder predictors.

Objective 2: Development of a novel protein-level predictor recommendation system to improve predictive performance of disorder predictions.

- Investigated the usefulness of physiochemical properties to predict expected predictive performance for individual proteins from selected set of computational disorder predictors. To the best of my knowledge, this is the first published study on this topic.
- Designed, developed and tested the machine learning (regression) based approach to predict expected predictive performance of a given protein for an individual disorder predictor.
- Designed, implemented and tested the DISOselect system that recommends the most accurate disorder predictor for a given input protein. To the best of my knowledge this is first such recommendation system for the computational disorder predictors.
- Developed and deployed the webserver that implements the DISOselect method.

Objective 3: Assessment and comparative analysis of the predictive performance of disorder predictions for specific functional types of disordered proteins.

- Collected and annotated the benchmark dataset.
- Created the similarity reduced benchmark dataset by clustering initial dataset with training datasets of evaluating predictors.
- Validated the experimental annotations for ordered regions by mapping them to a masked database of PDB sequences.
- Collected the disorder predictions from 10 different methods for the benchmark dataset.
- Evaluated the impact of similarity reduction of benchmark dataset to the predictive performance of the 10 selected methods.
- Evaluated the impact of the experimental validation of ordered regions on the predictive performance of the 10 selected methods.
- Evaluated the predictive performance of 10 selected methods for the functional subclasses of disordered proteins.

Objective 4: Accurate prediction of the disordered lipid-binding residues from protein sequences.

- Collected and functionally annotated the training and test datasets.
- Identified and investigated the usefulness of different physicochemical features to predict disordered lipid-binding residues.
- Designed and implemented DisoLipPred including the novel approach of using transfer learning and bypass strategy to predict the disordered lipid-binding residues from protein sequences.
- Collected the results and empirically assessed predictive performance of DisoLipPred against several approaches that can be used to indirectly predict the disordered lipid-binding residues.
- Evaluated the predictive performance of DisoLipPred on the benchmark test dataset, on the TM test dataset and on the complete yeast proteome.
- Developed and deployed a freely available webserver that implements the DisoLipPred method.

7.2 List of related publications

Journal:

1. **Katuwawala A**, Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions. *Briefings in Bioinformatics*. 2019
2. **Katuwawala A**, Oldfield CJ, Kurgan L. DISOselect: Disorder predictor selection at the protein level. *Protein Science*. 2020;29(1):184-200.
3. **Katuwawala A**, Peng Z, Yang J, Kurgan L. Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions. *Computational and Structural Biotechnology Journal*. 2019;17:454-62.
4. **Katuwawala, A.**; Kurgan, L. Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins. *Biomolecules* 2020, 10, 1636,

5. **Katuwawala, A.**, et al., DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server. *Journal of Molecular Biology*, 2019.
6. **Katuwawala, A.**, et al., QUARTERplus: accurate disorder predictions integrated with interpretable residue-level quality assessment scores. *Computational and Structural Biotechnology Journal*. 2021;
7. Zhao, B.; **Katuwawala, A.**; Uversky, V. N.; Kurgan, L., IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell. *Cellular and Molecular Life Sciences* 2020.
8. Zhao, B.; **Katuwawala, A.**; et al., DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Research*, 2020.

Book chapter:

9. **Katuwawala A**, Ghadermarzi S, Kurgan L. Chapter Nine - Computational prediction of functions of intrinsically disordered regions. In: Uversky VN, editor. *Progress in Molecular Biology and Translational Science*. 166: Academic Press; 2019. p. 341-69

Conference:

10. **Katuwawala, A.**; Ghadermarzi, S.; Oldfield, C. J.; Barik, A.; Kurgan, L., Disordered Function Conjunction: On the in-silico function annotation of intrinsically disordered regions. In *Biocomputing 2020*, pp 171-182.

References

1. Crick, F., *Central Dogma of Molecular Biology*. Nature, 1970. **227**(5258): p. 561-563.
2. Habchi, J., et al., *Introducing Protein Intrinsic Disorder*. Chemical Reviews, 2014. **114**(13): p. 6561-6588.
3. Xue, B., A.K. Dunker, and V.N. Uversky, *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life*. Journal of Biomolecular Structure and Dynamics, 2012. **30**(2): p. 137-149.
4. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life*. Cellular and Molecular Life Sciences, 2015. **72**(1): p. 137-151.
5. Dunker, A.K., et al., *Intrinsic Disorder and Protein Function*. Biochemistry, 2002. **41**(21): p. 6573-6582.
6. Dunker, A.K., et al., *Function and structure of inherently disordered proteins*. Current Opinion in Structural Biology, 2008. **18**(6): p. 756-764.
7. Xie, H., et al., *Functional Anthology of Intrinsic Disorder. 3. Ligands, Post-Translational Modifications, and Diseases Associated with Intrinsically Disordered Proteins*. Journal of Proteome Research, 2007. **6**(5): p. 1917-1932.
8. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020*. Nucleic Acids Research, 2019. **48**(D1): p. D269-D276.
9. Berman, H.M., et al., *The Protein Data Bank*. Nucleic acids research, 2000. **28**(1): p. 235-242.
10. Li, J., et al., *An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014*. International journal of molecular sciences, 2015. **16**(10): p. 23446-23462.
11. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions*. Cellular and Molecular Life Sciences, 2017. **74**(17): p. 3069-3090.
12. Atkins, J.D., et al., *Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies*. International journal of molecular sciences, 2015. **16**(8): p. 19040-19054.
13. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods*. Molecular bioSystems, 2012. **8**(1): p. 114-121.
14. Liu, Y., X. Wang, and B. Liu, *A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction*. Brief Bioinform, 2019. **20**(1): p. 330-346.
15. Zhen-Ling, P. and K. Lukasz, *Comprehensive Comparative Assessment of In-Silico Predictors of Disordered Regions*. Current Protein & Peptide Science, 2012. **13**(1): p. 6-18.
16. Dosztányi, Z., B. Mészáros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins*. Briefings in Bioinformatics, 2009. **11**(2): p. 225-243.
17. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10*. Proteins, 2014. **82 Suppl 2**(0 2): p. 127-137.
18. Necci, M., et al., *A comprehensive assessment of long intrinsic protein disorder from the DisProt database*. Bioinformatics, 2017. **34**(3): p. 445-452.
19. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Chapter Nine - Computational prediction of functions of intrinsically disordered regions*, in *Progress in Molecular Biology and Translational Science*, V.N. Uversky, Editor. 2019, Academic Press. p. 341-369.
20. Vickery, H.B., *The origin of the word protein*. The Yale journal of biology and medicine, 1950. **22**(5): p. 387-393.

21. Dunker, A.K., et al., *What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered*. *Intrinsically disordered proteins*, 2013. **1**(1): p. e24157-e24157.
22. van der Lee, R., et al., *Classification of Intrinsically Disordered Regions and Proteins*. *Chemical Reviews*, 2014. **114**(13): p. 6589-6631.
23. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions*, in *Intrinsically Disordered Proteins*, N. Salvi, Editor. 2019, Academic Press. p. 1-34.
24. Ward, J.J., et al., *Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life*. *Journal of Molecular Biology*, 2004. **337**(3): p. 635-645.
25. Tompa, P., *Intrinsically unstructured proteins*. *Trends in Biochemical Sciences*, 2002. **27**(10): p. 527-533.
26. Fukuchi, S., et al., *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners*. *Nucleic acids research*, 2014. **42**(Database issue): p. D320-D325.
27. Martin, A.J.M., I. Walsh, and S.C.E. Tosatto, *MOBI: a web server to define and visualize structural mobility in NMR protein ensembles*. *Bioinformatics*, 2010. **26**(22): p. 2916-2917.
28. Follis, A.V., et al., *The DNA-binding domain mediates both nuclear and cytosolic functions of p53*. *Nature Structural & Molecular Biology*, 2014. **21**(6): p. 535-543.
29. Follis, A.V., et al., *Regulation of apoptosis by an intrinsically disordered region of Bcl-xL*. *Nature chemical biology*, 2018. **14**(5): p. 458-465.
30. Muchmore, S.W., et al., *X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death*. *Nature*, 1996. **381**(6580): p. 335-341.
31. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
32. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics*. *Protein science : a publication of the Protein Society*, 2002. **11**(4): p. 739-756.
33. Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins*. *Nucleic Acids Res*, 2017. **45**(D1): p. D219-D227.
34. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020*. *Nucleic Acids Res*, 2020. **48**(D1): p. D269-D276.
35. Tskhovrebova, L. and J. Trinick, *Titin: properties and family relationships*. *Nature Reviews Molecular Cell Biology*, 2003. **4**(9): p. 679-689.
36. Collins, M.O., et al., *Phosphoproteomic Analysis of the Mouse Brain Cytosol Reveals a Predominance of Protein Phosphorylation in Regions of Intrinsic Sequence Disorder*. *Molecular & Cellular Proteomics*, 2008. **7**(7): p. 1331-1348.
37. Galea, C.A., et al., *Regulation of Cell Division by Intrinsically Unstructured Proteins: Intrinsic Flexibility, Modularity, and Signaling Conduits*. *Biochemistry*, 2008. **47**(29): p. 7598-7609.
38. Schroeder, R., A. Barta, and K. Semrad, *Strategies for RNA folding and assembly*. *Nature Reviews Molecular Cell Biology*, 2004. **5**(11): p. 908-919.
39. TOMPA, P. and P. CSERMELY, *The role of structural disorder in the function of RNA and protein chaperones*. *The FASEB Journal*, 2004. **18**(11): p. 1169-1175.
40. Sugase, K., H.J. Dyson, and P.E. Wright, *Mechanism of coupled folding and binding of an intrinsically disordered protein*. *Nature*, 2007. **447**(7147): p. 1021-1025.
41. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life*. *Molecular BioSystems*, 2016. **12**(3): p. 697-710.
42. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome*. *Cellular and Molecular Life Sciences*, 2014. **71**(8): p. 1477-1504.
43. Adilakshmi, T., P. Ramaswamy, and S.A. Woodson, *Protein-independent Folding Pathway of the 16S rRNA 5' Domain*. *Journal of Molecular Biology*, 2005. **351**(3): p. 508-519.

44. Xue, B., et al., *Stochastic machines as a colocalization mechanism for scaffold protein function*. FEBS Letters, 2013. **587**(11): p. 1587-1591.
45. Daniels, A.J., R.J.P. Williams, and P.E. Wright, *The character of the stored molecules in chromaffin granules of the adrenal medulla: A nuclear magnetic resonance study*. Neuroscience, 1978. **3**(6): p. 573-585.
46. Barik, A., et al., *DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server*. Journal of Molecular Biology, 2019.
47. Dunker, A.K., et al., *Intrinsically disordered protein*. Journal of Molecular Graphics and Modelling, 2001. **19**(1): p. 26-59.
48. Uversky, V.N., *What does it mean to be natively unfolded?* European Journal of Biochemistry, 2002. **269**(1): p. 2-12.
49. He, B., et al., *Predicting intrinsic disorder in proteins: an overview*. Cell Research, 2009. **19**(8): p. 929-949.
50. Liu, J. and B. Rost, *NORSp: predictions of long regions without regular secondary structure*. Nucleic Acids Research, 2003. **31**(13): p. 3833-3835.
51. Linding, R., et al., *GlobPlot: exploring protein sequences for globularity and disorder*. Nucleic Acids Research, 2003. **31**(13): p. 3701-3708.
52. Mészáros, B., G. Erdős, and Z. Dosztányi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding*. Nucleic Acids Research, 2018. **46**(W1): p. W329-W337.
53. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*. Bioinformatics, 2005. **21**(16): p. 3369-3376.
54. Linding, R., et al., *Protein Disorder Prediction: Implications for Structural Proteomics*. Structure, 2003. **11**(11): p. 1453-1459.
55. Cheng, J., M.J. Sweredoski, and P. Baldi, *Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data*. Data Mining and Knowledge Discovery, 2005. **11**(3): p. 213-222.
56. Ward, J.J., et al., *The DISOPRED server for the prediction of protein disorder*. Bioinformatics, 2004. **20**(13): p. 2138-2139.
57. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder*. BMC Bioinformatics, 2006. **7**(1): p. 208.
58. Zhang, T., et al., *SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method*. Journal of Biomolecular Structure and Dynamics, 2012. **29**(4): p. 799-813.
59. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks*. Bioinformatics, 2016. **33**(5): p. 685-692.
60. Fan, X. and L. Kurgan, *Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus*. Journal of Biomolecular Structure and Dynamics, 2014. **32**(3): p. 448-464.
61. Kozłowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins*. BMC Bioinformatics, 2012. **13**(1): p. 111.
62. Mizianty, M.J., Z.L. Peng, and L. Kurgan, *MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles*. Intrinsically Disordered Proteins, 2013. **1**(1): p. e24428.
63. Mizianty, M.J., V. Uversky, and L. Kurgan, *Prediction of intrinsic disorder in proteins using MFDp2*. Methods Mol Biol, 2014. **1137**: p. 147-62.
64. Walsh, I., et al., *CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs*. Nucleic Acids Research, 2011. **39**(suppl_2): p. W190-W196.

65. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder*. *Bioinformatics*, 2011. **28**(4): p. 503-509.
66. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins*. *Nucleic Acids Research*, 2017. **46**(D1): p. D471-D476.
67. Oates, M.E., et al., *D2P2: database of disordered protein predictions*. *Nucleic Acids Research*, 2012. **41**(D1): p. D508-D516.
68. Melamud, E. and J. Moult, *Evaluation of disorder predictions in CASP5*. *Proteins*, 2003. **53 Suppl 6**: p. 561-5.
69. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions*. *Brief Bioinform*, 2020. **21**(5): p. 1509-1522.
70. Necci, M., et al., *A comprehensive assessment of long intrinsic protein disorder from the DisProt database*. *Bioinformatics*, 2018. **34**(3): p. 445-452.
71. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder*. *Bioinformatics*, 2015. **31**(2): p. 201-8.
72. Pryor, E.E., Jr. and M.C. Wiener, *A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder*. *Biophys J*, 2014. **106**(8): p. 1638-49.
73. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10*. *Proteins*, 2014. **82 Suppl 2**: p. 127-37.
74. Peng, Z.L. and L. Kurgan, *Comprehensive comparative assessment of in-silico predictors of disordered regions*. *Curr Protein Pept Sci*, 2012. **13**(1): p. 6-18.
75. Monastyrskyy, B., et al., *Evaluation of disorder predictions in CASP9*. *Proteins*, 2011. **79 Suppl 10**: p. 107-18.
76. Noivirt-Brik, O., J. Prilusky, and J.L. Sussman, *Assessment of disorder predictions in CASP8*. *Proteins*, 2009. **77 Suppl 9**: p. 210-6.
77. Bordoli, L., F. Kiefer, and T. Schwede, *Assessment of disorder predictions in CASP7*. *Proteins*, 2007. **69 Suppl 8**: p. 129-36.
78. Jin, Y. and R.L. Dunbrack, Jr., *Assessment of disorder predictions in CASP6*. *Proteins*, 2005. **61 Suppl 7**: p. 167-75.
79. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10*. *Proteins: Structure, Function, and Bioinformatics*, 2014. **82**(S2): p. 127-137.
80. Fawcett, T., *An introduction to ROC analysis*. *Pattern Recognition Letters*, 2006. **27**(8): p. 861-874.
81. Nielsen, J.T. and F.A.A. Mulder, *Quality and bias of protein disorder predictors*. *Scientific Reports*, 2019. **9**(1): p. 5137.
82. Wang, S., et al., *DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields*. *International Journal of Molecular Sciences*, 2015. **16**(8): p. 17315-17330.
83. Oldfield, C.J., et al., *Comparing and Combining Predictors of Mostly Disordered Proteins†*. *Biochemistry*, 2005. **44**(6): p. 1989-2000.
84. Cheng, Y., et al., *Mining α -Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments†*. *Biochemistry*, 2007. **46**(47): p. 13468-13477.
85. Xue, B., A.K. Dunker, and V.N. Uversky, *Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction*. *International Journal of Molecular Sciences*, 2010. **11**(10): p. 3725-3747.
86. Oldfield, C.J., V.N. Uversky, and L. Kurgan, *Predicting functions of disordered proteins with MoRFPred*, in *Computational Methods in Protein Evolution*. 2019, Springer. p. 337-352.

87. Disfani, F.M., et al., *MoRFPred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins*. *Bioinformatics*, 2012. **28**(12): p. i75-i83.
88. Fang, C., et al., *MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation*. *BMC Bioinformatics*, 2013. **14**: p. 300.
89. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life*. *Mol Biosyst*, 2016. **12**(3): p. 697-710.
90. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity*. *Bioinformatics*, 2015. **31**(6): p. 857-63.
91. Malhis, N. and J. Gsponer, *Computational identification of MoRFs in protein sequences*. *Bioinformatics*, 2015. **31**(11): p. 1738-1744.
92. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences*. *Nucleic Acids Res*, 2016.
93. Sharma, R., et al., *Predicting MoRFs in protein sequences using HMM profiles*. *BMC Bioinformatics*, 2016. **17**.
94. Wang, H., et al., *Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity*. *Brief Bioinform*, 2018. **19**(5): p. 838-852.
95. Sharma, R., et al., *MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles*. *Journal of Theoretical Biology*, 2018. **437**: p. 9-16.
96. Sharma, R., et al., *OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences*. *Bioinformatics*, 2018. **34**: p. 1850-1858.
97. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences*. *Proteomics*, 2018. **1800058**: p. 1800058.
98. Fang, C., et al., *Identifying MoRFs in Disordered Proteins Using Enlarged Conserved Features*, in *ICBCB 2018 Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*. 2018. p. 50-54.
99. Sharma, R., et al., *Discovering MoRFs by trisecting intrinsically disordered protein sequence into terminals and middle regions*. *BMC bioinformatics*, 2019. **19**(13): p. 378.
100. Mooney, C., et al., *Prediction of Short Linear Protein Binding Regions*. *Journal of Molecular Biology*, 2012. **415**(1): p. 193-204.
101. Fang, C., et al. *Identifying Protein Short Linear Motifs by Position-Specific Scoring Matrix*. in *International Conference on Swarm Intelligence*. 2016. Springer.
102. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins*. *Bioinformatics*, 2009. **25**(20): p. 2745-6.
103. Mészáros, B., I. Simon, and Z. Dosztányi, *Prediction of Protein Binding Regions in Disordered Proteins*. *PLoS Comput Biol*, 2009. **5**(5): p. e1000376.
104. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind*. *Methods Mol Biol*, 2017. **1484**: p. 187-203.
105. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder*. *Nucleic Acids Res*, 2015. **43**(18): p. e121.
106. Mészáros, B., G. Erdős, and Z. Dosztányi, *IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding*. *Nucleic Acids Research*, 2018. **46**: p. W329-W337.
107. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences*. *Bioinformatics*, 2016. **32**(12): p. i341-i350.

108. Meng, F. and L. Kurgan, *High-throughput prediction of disordered moonlighting regions in protein sequences*. *Proteins: Structure, Function, and Bioinformatics*, 2018. **86**(10): p. 1097-1110.
109. Mohan, A., et al., *Analysis of Molecular Recognition Features (MoRFs)*. *Journal of Molecular Biology*, 2006. **362**(5): p. 1043-1059.
110. Oldfield, C.J., et al., *Comparing and Combining Predictors of Mostly Disordered Proteins*. *Biochemistry*, 2005. **44**(6): p. 1989-2000.
111. Cheng, Y., et al., *Mining α -Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments*. *Biochemistry*, 2007. **46**(47): p. 13468-13477.
112. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity*. *Bioinformatics*, 2014. **31**(6): p. 857-863.
113. Van Roey, K., et al., *Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation*. *Chemical Reviews*, 2014. **114**(13): p. 6733-6778.
114. Fang, C., et al. *Identifying Protein Short Linear Motifs by Position-Specific Scoring Matrix*. 2016. Cham: Springer International Publishing.
115. Dosztányi, Z., B. Mészáros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins*. *Bioinformatics*, 2009. **25**(20): p. 2745-2746.
116. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind*, in *Prediction of Protein Secondary Structure*, Y. Zhou, et al., Editors. 2017, Springer New York: New York, NY. p. 187-203.
117. Schaefer, R.L., L.D. Roi, and R.A. Wolfe, *A ridge logistic estimator*. *Communications in Statistics - Theory and Methods*, 1984. **13**(1): p. 99-113.
118. Salzberg, S.L., *C4.5: Programs for Machine Learning by J. Ross Quinlan*. *Morgan Kaufmann Publishers, Inc., 1993*. *Machine Learning*, 1994. **16**(3): p. 235-240.
119. John, G.H. and P. Langley, *Estimating continuous distributions in Bayesian classifiers*, in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 1995, Morgan Kaufmann Publishers Inc.: Montréal, Qué, Canada. p. 338-345.
120. Breiman, L., *Random Forests*. *Machine Learning*, 2001. **45**(1): p. 5-32.
121. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning*. *Nature*, 2015. **521**(7553): p. 436-44.
122. Student, *The probable error of a mean*. *Biometrika*, 1908. **6**(1): p. 1-25.
123. Wilcoxon, F., *Individual comparisons by ranking methods*. *Biometrics bulletin*, 1945: p. 80-83.
124. Anderson, T.W. and D.A. Darling, *Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes*. *Ann. Math. Statist.*, 1952: p. 193-212.
125. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions*. *Briefings in Bioinformatics*, 2019.
126. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder*. *Bioinformatics*, 2014. **31**(2): p. 201-208.
127. Pryor, Edward E. and Michael C. Wiener, *A Critical Evaluation of in silico Methods for Detection of Membrane Protein Intrinsic Disorder*. *Biophysical Journal*, 2014. **106**(8): p. 1638-1649.
128. Liu, Y., X. Wang, and B. Liu, *A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction*. *Briefings in Bioinformatics*, 2017. **20**(1): p. 330-346.
129. Mizianty, M.J., V. Uversky, and L. Kurgan, *Prediction of Intrinsic Disorder in Proteins Using MFDp2*, in *Protein Structure Prediction*, D. Kihara, Editor. 2014, Springer New York: New York, NY. p. 147-162.
130. Postel, S., et al., *Bacterial flagellar capping proteins adopt diverse oligomeric states*. *eLife*, 2016. **5**: p. e18857.

131. Vo, A.T., et al., *Defining the domains of Cia2 required for its essential function in vivo and in vitro*. Metallomics, 2017. **9**(11): p. 1645-1654.
132. Jain, G., et al., *A Model Sea Urchin Spicule Matrix Protein Self-Associates To Form Mineral-Modifying Protein Hydrogels*. Biochemistry, 2016. **55**(31): p. 4410-4421.
133. Chang, E.P., et al., *Insect Cell Glycosylation and Its Impact on the Functionality of a Recombinant Intracrystalline Nacre Protein, AP24*. Biochemistry, 2016. **55**(7): p. 1024-1035.
134. Yadav, L.R., et al., *Functional assessment of intrinsic disorder central domains of BRCA1*. Journal of Biomolecular Structure and Dynamics, 2015. **33**(11): p. 2469-2478.
135. Camacho, C., et al., *BLAST+: architecture and applications*. BMC Bioinformatics, 2009. **10**(1): p. 421.
136. Hu, G., et al., *Quality assessment for the putative intrinsic disorder in proteins*. Bioinformatics, 2018. **35**(10): p. 1692-1700.
137. Wootton, J.C., *Non-globular domains in protein sequences: Automated segmentation using complexity measures*. Computers & Chemistry, 1994. **18**(3): p. 269-285.
138. Jones, D.T. and M.B. Swindells, *Getting the most from PSI-BLAST*. Trends in Biochemical Sciences, 2002. **27**(3): p. 161-164.
139. Prilusky, J., et al., *FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded*. Bioinformatics, 2005. **21**(16): p. 3435-3438.
140. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *DISOselect: Disorder predictor selection at the protein level*. Protein Science, 2020. **29**(1): p. 184-200.
141. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins: Structure, Function, and Bioinformatics, 2000. **41**(3): p. 415-427.
142. Andrew, C., et al., *TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder*. Protein & Peptide Letters, 2008. **15**(9): p. 956-963.
143. Szilágyi, A., D. Györfy, and P. Závodszy, *The Twilight Zone between Protein Order and Disorder*. Biophysical Journal, 2008. **95**(4): p. 1612-1626.
144. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding*. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2010. **1804**(6): p. 1231-1264.
145. Uversky, V.N., *Intrinsically disordered proteins from A to Z*. The International Journal of Biochemistry & Cell Biology, 2011. **43**(8): p. 1090-1103.
146. Faraggi, E., Y. Zhou, and A. Kloczkowski, *Accurate single-sequence prediction of solvent accessible surface area using local and global features*. Proteins: Structure, Function, and Bioinformatics, 2014. **82**(11): p. 3170-3176.
147. Buchan, D.W.A., et al., *Scalable web services for the PSIPRED Protein Analysis Workbench*. Nucleic Acids Research, 2013. **41**(W1): p. W349-W357.
148. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008*. Nucleic acids research, 2008. **36**(Database issue): p. D202-D205.
149. Altman, N.S., *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*. The American Statistician, 1992. **46**(3): p. 175-185.
150. Freedman, D.A., *Statistical Models : Theory and Practice*. 2009, New York, UNITED STATES: Cambridge University Press.
151. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees*. Machine Learning, 2006. **63**(1): p. 3-42.
152. Breiman, L., *Randomizing Outputs to Increase Prediction Accuracy*. Machine Learning, 2000. **40**(3): p. 229-242.

153. Quinlan, J.R., *DECISION TREES AS PROBABILISTIC CLASSIFIERS*, in *Proceedings of the Fourth International Workshop on MACHINE LEARNING*, P. Langley, Editor. 1987, Morgan Kaufmann. p. 31-37.
154. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines*. *Nucleic Acids Research*, 2006. **34**(suppl_2): p. W164-W168.
155. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. *Nucleic Acids Research*, 1997. **25**(17): p. 3389-3402.
156. McGinnis, S. and T.L. Madden, *BLAST: at the core of a powerful and diverse set of sequence analysis tools*. *Nucleic Acids Research*, 2004. **32**(suppl_2): p. W20-W25.
157. Punta, M., et al., *The Pfam protein families database*. *Nucleic Acids Research*, 2011. **40**(D1): p. D290-D301.
158. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments*. *Nucleic Acids Research*, 2007. **36**(suppl_1): p. D419-D425.
159. de Lima Morais, D.A., et al., *SUPERFAMILY 1.75 including a domain-centric gene ontology method*. *Nucleic Acids Research*, 2010. **39**(suppl_1): p. D427-D434.
160. Sillitoe, I., et al., *New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures*. *Nucleic Acids Research*, 2012. **41**(D1): p. D490-D498.
161. Gsponer, J. and M. Madan Babu, *The rules of disorder or why disorder rules*. *Progress in Biophysics and Molecular Biology*, 2009. **99**(2): p. 94-103.
162. Sickmeier, M., et al., *DisProt: the Database of Disordered Proteins*. *Nucleic Acids Res*, 2007. **35**(Database issue): p. D786-93.
163. Dinkel, H., et al., *The eukaryotic linear motif resource ELM: 10 years and counting*. *Nucleic Acids Research*, 2013. **42**(D1): p. D259-D266.
164. Mao, A.H., et al., *Net charge per residue modulates conformational ensembles of intrinsically disordered proteins*. *Proceedings of the National Academy of Sciences*, 2010. **107**(18): p. 8183-8188.
165. Das, R.K. and R.V. Pappu, *Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues*. *Proceedings of the National Academy of Sciences*, 2013. **110**(33): p. 13392-13397.
166. Katuwawala, A. and L. Kurgan, *Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins*. *Biomolecules*, 2020. **10**(12).
167. Wang, C., V.N. Uversky, and L. Kurgan, *Disordered nucleome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea*. *Proteomics*, 2016. **16**(10): p. 1486-98.
168. Meng, F., et al., *Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments*. *Int J Mol Sci*, 2015. **17**(1).
169. Dyson, H.J., *Roles of intrinsic disorder in protein-nucleic acid interactions*. *Mol Biosyst*, 2012. **8**(1): p. 97-104.
170. Dunker, A.K., et al., *Flexible nets. The roles of intrinsic disorder in protein interaction networks*. *FEBS J*, 2005. **272**(20): p. 5129-48.
171. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Computational prediction of functions of intrinsically disordered regions*. *Prog Mol Biol Transl Sci*, 2019. **166**: p. 341-369.
172. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions*. *Comput Struct Biotechnol J*, 2019. **17**: p. 454-462.
173. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions*. *Cell Mol Life Sci*, 2017. **74**(17): p. 3069-3090.
174. Meng, F., V. Uversky, and L. Kurgan, *Computational Prediction of Intrinsic Disorder in Proteins*. *Curr Protoc Protein Sci*, 2017. **88**: p. 2 16 1-2 16 14.

175. Uversky, V.N., *How to Predict Disorder in a Protein of Interest*. Methods Mol Biol, 2017. **1484**: p. 137-158.
176. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe*. Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
177. Li, J., et al., *An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014*. Int J Mol Sci, 2015. **16**(10): p. 23446-62.
178. Deng, X., et al., *An Overview of Practical Applications of Protein Disorder Prediction and Drive for Faster, More Accurate Predictions*. Int J Mol Sci, 2015. **16**(7): p. 15384-404.
179. Bhowmick, P., M. Guharoy, and P. Tompa, *Bioinformatics Approaches for Predicting Disordered Protein Motifs*. Adv Exp Med Biol, 2015. **870**: p. 291-318.
180. Varadi, M., et al., *Computational approaches for inferring the functions of intrinsically disordered proteins*. Front Mol Biosci, 2015. **2**: p. 45.
181. Atkins, J.D., et al., *Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies*. Int J Mol Sci, 2015. **16**(8): p. 19040-54.
182. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods*. Mol Biosyst, 2012. **8**(1): p. 114-21.
183. Orosz, F. and J. Ovadi, *Proteins without 3D structure: definition, detection and beyond*. Bioinformatics, 2011. **27**(11): p. 1449-54.
184. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins*. Brief Bioinform, 2010. **11**(2): p. 225-43.
185. He, B., et al., *Predicting intrinsic disorder in proteins: an overview*. Cell Res, 2009. **19**(8): p. 929-49.
186. Uversky, V.N., et al., *Prediction of intrinsic disorder and its use in functional proteomics*. Methods Mol Biol, 2007. **408**: p. 69-92.
187. van der Lee, R., et al., *Classification of intrinsically disordered regions and proteins*. Chem Rev, 2014. **114**(13): p. 6589-631.
188. Dunker, A.K., et al., *Intrinsic disorder and protein function*. Biochemistry, 2002. **41**(21): p. 6573-82.
189. Tompa, P., *Intrinsically unstructured proteins*. Trends Biochem Sci, 2002. **27**(10): p. 527-533.
190. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks*. Bioinformatics, 2017. **33**(5): p. 685-692.
191. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity*. Bioinformatics, 2015. **31**(6): p. 857-863.
192. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder*. Bioinformatics, 2012. **28**(4): p. 503-509.
193. Dosztányi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content*. Bioinformatics, 2005. **21**(16): p. 3433-3434.
194. Dosztányi, Z., et al., *The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins*. Journal of Molecular Biology, 2005. **347**(4): p. 827-839.
195. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding*. Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.
196. Wang, S., J.Z. Ma, and J.B. Xu, *AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields*. Bioinformatics, 2016. **32**(17): p. 672-679.
197. Wang, S., et al., *DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields*. Int J Mol Sci, 2015. **16**(8): p. 17315-30.

198. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks*. *Bioinformatics*, 2017. **33**(5): p. 685-692.
199. Hanson, J., K. Paliwal, and Y. Zhou, *Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures*. *J Chem Inf Model*, 2018. **58**(11): p. 2369-2376.
200. Kedarisetti, K.D., et al., *Improved sequence-based prediction of strand residues*. *J Bioinform Comput Biol*, 2011. **9**(1): p. 67-89.
201. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data*. *Bioinformatics*, 2012. **28**(23): p. 3150-2.
202. Dana, J.M., et al., *SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins*. *Nucleic Acids Res*, 2019. **47**(D1): p. D482-D489.
203. Hanson, J., et al., *SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning*. *Genomics Proteomics Bioinformatics*, 2020.
204. Zhang, J., Z. Ma, and L. Kurgan, *Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains*. *Brief Bioinform*, 2019. **20**(4): p. 1250-1268.
205. Necci, M., D. Piovesan, and S.C.E. Tosatto, *Critical Assessment of Protein Intrinsic Disorder Prediction*. *bioRxiv*, 2020: p. 2020.08.11.245852.
206. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Computational prediction of functions of intrinsically disordered regions*, in *Progress in Molecular Biology and Translational Science*, V.N. Uversky, Editor. 2019, Academic Press. p. 341-369.
207. Fuxreiter, M., et al., *Disordered proteinaceous machines*. *Chem Rev*, 2014. **114**(13): p. 6806-43.
208. Balcerak, A., et al., *RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity*. *Open Biol*, 2019. **9**(6): p. 190096.
209. Varadi, M., et al., *Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins*. *PLoS One*, 2015. **10**(10): p. e0139731.
210. Patil, A., K. Kinoshita, and H. Nakamura, *Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network*. *Protein Sci*, 2010. **19**(8): p. 1461-8.
211. Kjaergaard, M. and B.B. Kragelund, *Functions of intrinsic disorder in transmembrane proteins*. *Cellular and Molecular Life Sciences*, 2017. **74**(17): p. 3205-3224.
212. Ghadermarzi, S., et al., *Disordered Function Conjunction: On the in-silico function annotation of intrinsically disordered regions*. *Pac Symp Biocomput*, 2020. **25**: p. 171-182.
213. Soto-Avellaneda, A. and B.E. Morrison, *Signaling and other functions of lipids in autophagy: a review*. *Lipids in Health and Disease*, 2020. **19**(1): p. 214.
214. Di Paolo, G. and P. De Camilli, *Phosphoinositides in cell regulation and membrane dynamics*. *Nature*, 2006. **443**(7112): p. 651-657.
215. Dall'Armi, C., Kelly A. Devereaux, and G. Di Paolo, *The Role of Lipids in the Control of Autophagy*. *Current Biology*, 2013. **23**(1): p. R33-R45.
216. Settembre, C., et al., *TFEB controls cellular lipid metabolism through a starvation-induced autoregulatory loop*. *Nature Cell Biology*, 2013. **15**(6): p. 647-658.
217. and, D.A.B. and E. London, *FUNCTIONS OF LIPID RAFTS IN BIOLOGICAL MEMBRANES*. *Annual Review of Cell and Developmental Biology*, 1998. **14**(1): p. 111-136.
218. Welte, M.A. and A.P. Gould, *Lipid droplet functions beyond energy storage*. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 2017. **1862**(10, Part B): p. 1260-1272.
219. Gorbenko, G.P. and P.K.J. Kinnunen, *The role of lipid-protein interactions in amyloid-type protein fibril formation*. *Chemistry and Physics of Lipids*, 2006. **141**(1): p. 72-82.

220. Han, X. and L.K. Tamm, *pH-dependent Self-association of Influenza Hemagglutinin Fusion Peptides in Lipid Bilayers*. Journal of Molecular Biology, 2000. **304**(5): p. 953-965.
221. Bokvist, M., et al., *Two Types of Alzheimer's β -Amyloid (1–40) Peptide Membrane Interactions: Aggregation Preventing Transmembrane Anchoring Versus Accelerated Surface Fibril Formation*. Journal of Molecular Biology, 2004. **335**(4): p. 1039-1049.
222. Necula, M., C.N. Chirita, and J. Kuret, *Rapid Anionic Micelle-mediated α -Synuclein Fibrillization in Vitro**. Journal of Biological Chemistry, 2003. **278**(47): p. 46674-46680.
223. Knyazeva, E.L., et al., *Who Is Mr. HAMLET? Interaction of Human α -Lactalbumin with Monomeric Oleic Acid*. Biochemistry, 2008. **47**(49): p. 13127-13137.
224. Assayag, K., et al., *Polyunsaturated Fatty Acids Induce α -Synuclein-Related Pathogenic Changes in Neuronal Cells*. The American Journal of Pathology, 2007. **171**(6): p. 2000-2011.
225. Chirita, C.N., M. Necula, and J. Kuret, *Anionic Micelles and Vesicles Induce Tau Fibrillization in Vitro **. Journal of Biological Chemistry, 2003. **278**(28): p. 25644-25650.
226. Chan, C.S., et al., *Towards understanding the Tat translocation mechanism through structural and biophysical studies of the amphipathic region of TatA from Escherichia coli*. Biochimica et Biophysica Acta (BBA) - Biomembranes, 2011. **1808**(9): p. 2289-2296.
227. Deryusheva, E., et al., *Does Intrinsic Disorder in Proteins Favor Their Interaction with Lipids?* PROTEOMICS, 2019. **19**(6): p. 1800098.
228. Ugalde, C.L., et al., *The role of lipids in α -synuclein misfolding and neurotoxicity*. The Journal of biological chemistry, 2019. **294**(23): p. 9016-9028.
229. Jesus, A., et al., *Assembly In Vitro of Tau Protein and its Implications in Alzheimers Disease*. Current Alzheimer Research, 2004. **1**(2): p. 97-101.
230. Patil, S. and C. Chan, *Palmitic and stearic fatty acids induce Alzheimer-like hyperphosphorylation of tau in primary rat cortical neurons*. Neuroscience Letters, 2005. **384**(3): p. 288-293.
231. Ruipérez, V., F. Darios, and B. Davletov, *Alpha-synuclein, lipids and Parkinson's disease*. Progress in Lipid Research, 2010. **49**(4): p. 420-428.
232. Song, M. and H. Kim, *Stability and Solvent Accessibility of SecA Protein of Escherichia coli1*. The Journal of Biochemistry, 1997. **122**(5): p. 1010-1018.
233. van der Goot, F.G., et al., *A 'molten-globule' membrane-insertion intermediate of the pore-forming domain of colicin A*. Nature, 1991. **354**(6352): p. 408-410.
234. Peters, C., et al., *Improved topology prediction using the terminal hydrophobic helices rule*. Bioinformatics, 2015. **32**(8): p. 1158-1162.
235. Käll, L., A. Krogh, and E.L.L. Sonnhammer, *Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server*. Nucleic Acids Research, 2007. **35**(suppl_2): p. W429-W432.
236. Roy Choudhury, A. and M. Novic, *PredbetaTM: A Novel beta-Transmembrane Region Prediction Algorithm*. PLoS One, 2015. **10**(12): p. e0145564.
237. Disfani, F.M., et al., *MoRFPred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins*. Bioinformatics, 2012. **28**(12): p. i75-83.
238. Malhis, N. and J. Gsponer, *Computational identification of MoRFs in protein sequences*. Bioinformatics, 2015. **31**(11): p. 1738-44.
239. Sharma, R., et al., *OPAL: prediction of MoRF regions in intrinsically disordered protein sequences*. Bioinformatics, 2018. **34**(11): p. 1850-1858.
240. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences*. Proteomics, 2019. **19**(6): p. e1800058.
241. wwPDB consortium, *Protein Data Bank: the single global archive for 3D macromolecular structure data*. Nucleic Acids Res, 2019. **47**(D1): p. D520-D528.

242. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences*. Bioinformatics, 2010. **26**(5): p. 680-682.
243. Faraggi, E., Y.Q. Zhou, and A. Kloczkowski, *Accurate single-sequence prediction of solvent accessible surface area using local and global features*. Proteins, 2014. **82**(11): p. 3170-3176.
244. Huang, H.L., et al., *Predicting and Analyzing Lipid-Binding Proteins Using an Efficient Physicochemical Property Mining Method*. Applied Mechanics and Materials, 2013. **421**: p. 313-318.
245. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008*. Nucleic Acids Res, 2008. **36**(Database issue): p. D202-5.
246. Weiss, K., T.M. Khoshgoftaar, and D. Wang, *A survey of transfer learning*. Journal of Big Data, 2016. **3**(1): p. 9.
247. Tan, C., et al. *A Survey on Deep Transfer Learning*. 2018. Cham: Springer International Publishing.
248. Singh, J., et al., *RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning*. Nature Communications, 2019. **10**.
249. Li, F., et al., *DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites*. Bioinformatics, 2020. **36**(4): p. 1057-1065.
250. Jin, J., et al., *Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism*. Proteins, 2021.
251. Liu, L., et al., *TSPTFBS: a docker image for Trans-Species Prediction of Transcription Factor Binding Sites in Plants*. Bioinformatics, 2021.
252. Fang, C., et al., *Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method*. J Bioinform Comput Biol, 2019. **17**(1): p. 1950004.
253. Hanson, J., et al., *Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning*. Bioinformatics, 2020. **36**(4): p. 1107-1113.
254. UniProt, C., *UniProt: the universal protein knowledgebase in 2021*. Nucleic Acids Res, 2021. **49**(D1): p. D480-D489.
255. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs*. Bioinformatics, 2015. **31**(19): p. 3210-3212.
256. Ize, B., et al., *In vivo dissection of the Tat translocation pathway in Escherichia coli* Edited by G. von Heijne. Journal of Molecular Biology, 2002. **317**(3): p. 327-335.
257. Sargent, F., et al., *Overlapping functions of components of a bacterial Sec-independent protein export pathway*. The EMBO journal, 1998. **17**(13): p. 3640-3650.
258. Zhang, Y., et al., *Structural Basis for Tata Oligomerization: An NMR Study of Escherichia coli Tata Dimeric Structure*. PLOS ONE, 2014. **9**(8): p. e103157.
259. Porcelli, I., et al., *Characterization and Membrane Assembly of the Tata Component of the Escherichia coli Twin-Arginine Protein Transport System*. Biochemistry, 2002. **41**(46): p. 13690-13697.
260. Faraggi, E., Y. Zhou, and A. Kloczkowski, *Accurate single-sequence prediction of solvent accessible surface area using local and global features*. Proteins, 2014. **82**(11): p. 3170-3176.

Glossary

This section briefly describes the key terminology used in this document (in alphabetical order).

Disorder Function: A biological or biochemical function that is assigned to an intrinsically disordered region or intrinsically disordered protein.

Disorder Predictor: A computational method which predict disorder for individual amino acids in the protein sequence.

Disordered Protein: A Protein that includes one or more intrinsically disordered regions.

Disordered Region: A unique amino acid sequence that have no single, well-defined equilibrium structure and exist as highly dynamic, heterogeneous ensembles of conformers resulting from their relatively flat free energy surface.

Individual Protein Level: A complete protein sequence. Individual proteins usually have unique identifiers that rely on a well-defined naming convention.

Protein Dataset: A collection of multiple proteins, typically used to perform assessment of disorder predictors.

Protein Region Level: A sequence of multiple consecutive residues in a single protein.

Protein Residue Level: The smallest building unit of proteins. This unit is also referred to as an amino acid.

Appendix 1 – Complete set of 130 features used to implement the disorder predictor recommendation system (the DISOselect method)

Category	Feature name	Description	Source
Amino Acid composition (20 features)	Alanine Content	Fraction of Alanine in the input protein chain	Input sequence
	Leucine Content	Fraction of Leucine in the input protein chain	Input sequence
	Arginine Content	Fraction of Arginine in the input protein chain	Input sequence
	Asparagine Content	Fraction of Asparagine in the input protein chain	Input sequence
	Aspartic Content	Fraction of Aspartic acid in the input protein chain	Input sequence
	Cysteine Content	Fraction of Cysteine in the input protein chain	Input sequence
	Glutamic Content	Fraction of Glutamic acid in the input protein chain	Input sequence
	Glutamine Content	Fraction of Glutamine in the input protein chain	Input sequence
	Glycine Content	Fraction of Glycine in the input protein chain	Input sequence
	Histidine Content	Fraction of Histidine in the input protein chain	Input sequence
	Isoleucine Content	Fraction of Isoleucine in the input protein chain	Input sequence
	Lysine Content	Fraction of Lysine in the input protein chain	Input sequence
	Methionine Content	Fraction of Methionine in the input protein chain	Input sequence
	Phenylalanine Content	Fraction of Phenylalanine in the input protein chain	Input sequence
	Proline Content	Fraction of Proline in the input protein chain	Input sequence
	Serine Content	Fraction of Serine in the input protein chain	Input sequence
	Threonine Content	Fraction of Threonine in the input protein chain	Input sequence
	Tryptophan Content	Fraction of Tryptophan in the input protein chain	Input sequence
	Tyrosine Content	Fraction of Tyrosine in the input protein chain	Input sequence
	Valine Content	Fraction of Valine in the input protein chain	Input sequence
Predicted Solvent Accessibility (3 features)	Total accessible surface area	Sum of solvent accessibility of all residues	Predicted with ASAquick [260]
	Average accessible surface area	Average of solvent accessibility of all residues	Predicted with ASAquick [260]
	Total number of exposed residues	Sum of binary exposed residues	Predicted with ASAquick [260]
Sequence Complexity (2 features)	Fraction of complex regions	Number of complex regions divided by chain length	Computed by SEG [137]
	Fraction of complex residues	Number of complex residues divided by chain length	Computed by SEG [137]
Predicted Secondary Structure	Count of coils	Count of putative coil residues in protein	Predicted with PSIPRED [147]
	Count of helices	Count of putative helix residues in protein	Predicted with PSIPRED [147]

(8 features)	Count of strands	Count of putative strand residues in protein	Predicted with PSIPRED [147]
	Count of coils and strands	Count of putative coil and strand residues in protein	Predicted with PSIPRED [147]
	Content of coils	Fraction of putative coil residues in the input protein chain	Predicted with PSIPRED [147]
	Content of helices	Fraction of putative helix residues in the input protein chain	Predicted with PSIPRED [147]
	Content of strands	Fraction of putative strands residues in the input protein chain	Predicted with PSIPRED [147]
	Content of coils and strands	Fraction of putative coils and strand residues in the input protein chain	Predicted with PSIPRED [147]
Physiochemical properties of amino acids (97 features)	Summed hydropathy	Sum of hydropathy values of all residues	Extracted from AAindex [245]: KYTJ820101
	Summed net charge	Sum of net charge values of all residues	Extracted from AAindex [245]: KLEP840101
	Summed hydrophilicity	Sum of hydrophilicity values of all residues of all residues	Extracted from AAindex [245]: HOPT810101
	Average hydrophilicity	Sum of hydrophilicity values divided by chain length	Extracted from AAindex [245]: HOPT810101
	Average absolute entropy	Sum of absolute entropy values divided by chain length	Extracted from AAindex [245]: HUTJ700102
	Average unfolding gibbs energy	Sum of unfolding Gibbs energy values divided by chain length	Extracted from AAindex [245]: YUTK870101
	Average beta coils	Sum of beta-structure-coil equilibrium constants divided by chain length	Extracted from AAindex [245]: OOBM850101
	Average reverse turns	Sum of propensities to form reverse turn divided by chain length	Extracted from AAindex [4]: OOBM850102
	Summed transfer energy	Sum of transfer energy parameters of all residues	Extracted from AAindex [4]: OOBM850103
	Average Isoelectricity	Sum of isoelectric points divided by chain length	Extracted from AAindex [4]: ZIMJ680104
	Sequence complexity	Sum of composite amino acid of all residues	Raw sequence
	Summed hydrophobicity	Sum of hydrophobicity values of all residues	Extracted from AAindex [4]: PRAM900101
	Average hydrophobicity	Sum of hydrophobicity values divided by chain length	Extracted from AAindex [4]: PRAM900101
	Average hydropathy	Sum of hydropathy values divided by chain length	Extracted from AAindex [4]: KYTJ820101
	Summed solvation free energy	Sum of solvation free energy values of all residues	Extracted from AAindex [4]: EISD860101
	Average solvation free energy	Sum of solvation free energy values divided by chain length	Extracted from AAindex [4]: EISD860101
	Summed polarity	Sum of polarity values of all residues	Extracted from AAindex [4]: GRAR740102
	Average polarity	Sum of polarity values divided by chain length	Extracted from AAindex [4]: GRAR740102
	Summed volume	Sum of volume values of all residues	Extracted from AAindex [4]: GRAR740103
	Average volume	Sum of volume values divided by chain length	Extracted from AAindex [4]: GRAR740103
	Summed absolute entropy	Sum of absolute entropy values of all residues	Extracted from AAindex [4]: HUTJ700102
	Summed unfolding gibbs	Sum of unfolding Gibbs energy in water of all residues	Extracted from AAindex [4]: YUTK870101
	Summed activation gibbs	Sum of activation Gibbs energy values of all residues	Extracted from AAindex [4]: KLEP840101
	Average activation gibbs	Sum of activation Gibbs energy values divided by chain length	Extracted from AAindex [4]: KLEP840101
	Summed beta coils	Sum of beta-structure-coil equilibrium constants of all residues	Extracted from AAindex [4]: OOBM850101
	Summed reverse turn	Sum of propensity to form reverse turn values of all residues	Extracted from AAindex [4]: OOBM850102
Average transfer energy	Sum of transfer energy parameters divided by chain length	Extracted from AAindex [4]: OOBM850103	

Summed isoelectric points	Sum of isoelectric points values of all residues	Extracted from AAindex [4]: ZIMJ680104
Summed charge transfer	Sum of parameter of charge transfer capability of all residues	Extracted from AAindex [4]: CHAM830107
Summed charge donor	Sum of parameter of charge donor capability of all residues	Extracted from AAindex [4]: CHAM830108
Summed positive charge	Sum of parameter of positive charge capability of all residues	Extracted from AAindex [4]: CHAM830108
Summed negative charge	Sum of parameter of negative charge capability of all residues	Extracted from AAindex [4]: CHAM830108
Summed hydrophobicity index	Sum of hydrophobicity Index values of all residues	Extracted from AAindex [4]: ARGP820101
Average hydrophobicity index	Sum of hydrophobicity Index values divided by chain length	Extracted from AAindex [4]: ARGP820101
Summed alpha hydrophobicity	Sum of normalized hydrophobicity scales for alpha-proteins of all residues	Extracted from AAindex [4]: CIDH920101
Average alpha hydrophobicity	Sum of normalized hydrophobicity scales for alpha-proteins divided by chain length	Extracted from AAindex [4]: CIDH920101
Summed normalized average hydrophobicity	Sum of normalized average hydrophobicity scales of all residues	Extracted from AAindex [4]: CIDH920105
Average normalized average hydrophobicity	Sum of Normalized average hydrophobicity scales divided by chain length	Extracted from AAindex [4]: CIDH920105
Summed consensus normalized hydrophobicity	Sum of consensus normalized hydrophobicity scales of all residues	Extracted from AAindex [4]: EISD840101
Average consensus normalized hydrophobicity	Sum of Consensus normalized hydrophobicity scales divided by chain length	Extracted from AAindex [4]: EISD840101
Summed average surrounding hydrophobicity	Sum of average surrounding hydrophobicity values of all residues	Extracted from AAindex [4]: MANP780101
Average surrounding hydrophobicity	Sum of average surrounding hydrophobicity values divided by chain length	Extracted from AAindex [4]: MANP780101
Summed hydrophobicityPH3	Sum of hydrophobicity index values at 3.0 pH of all residues	Extracted from AAindex [4]: COWR900101
Average hydrophobicityPH3	Sum of hydrophobicity index values at 3.0 pH divided by chain length	Extracted from AAindex [4]: COWR900101
Summed native hydrophobicity	Sum of native hydrophobicity index values of all residues	Extracted from AAindex [4]: CASG920101
Average native hydrophobicity	Sum of native Hydrophobicity index values divided by chain length	Extracted from AAindex [4]: CASG920101
Disorder complexity	Fraction of disorder promoting amino acids in the input protein chain	Input sequence
Order complexity	Fraction of order promoting amino acids in the input protein chain	Input sequence
Charge to hydropathy ratio	Total charge of a protein as ratio of total hydropathy of all residues	Calculated AA index
Disorder complexity to order complexity ratio	Ratio between disorder promoting amino acids fraction and order promoting amino acids fraction in the input protein chain	Input sequence
Summed mass	Sum of masses of all residues	Input sequence
Average mass	Sum of masses divided by chain length	Input sequence
Summed density	Total mass of a protein as a ratio of total volume of all residues	Calculated AA index
Average density	Total density of protein divided by chain length	Calculated AA index
Length of each protein	Number of amino acids in the input protein chain	Input sequence
Summed CH chemical shifts	Sum of alphaCH chemical shift values of all residues	Extracted from AAindex [4]: ANDN920101
Average CH chemical shifts	Sum of alphaCH chemical shift values divided by chain length	Extracted from AAindex [4]: ANDN920101

Summed NH chemical shifts	Sum of alphaNH chemical shift values of all residues	Extracted from AAindex [4]: BUNA790101
Average NH chemical shifts	Sum of alphaNH chemical shift values divided by chain length	Extracted from AAindex [4]: BUNA790101
Summed spin coupling	Sum of spin coupling constants of all residues	Extracted from AAindex [4]: BUNA790103
Average spin coupling	Sum of spin coupling constants divided by chain length	Extracted from AAindex [4]: BUNA790103
Summed membrane preference	Sum of membrane preference indexes of all residues	Extracted from AAindex [4]: DESM900101
Average membrane preference	Sum of membrane preference indexes divided by chain length	Extracted from AAindex [4]: DESM900101
Summed hydrophobic moment	Sum of atom-based hydrophobic moment values of all residues	Extracted from AAindex [4]: EISD860102
Average hydrophobic moment	Sum of atom-based hydrophobic moment values divided by chain length	Extracted from AAindex [4]: EISD860102
Summed hydrophobic moment direction	Sum of direction of hydrophobic moment values of all residues	Extracted from AAindex [4]: EISD860103
Average hydrophobic moment direction	Sum of direction of hydrophobic moment values divided by chain length	Extracted from AAindex [4]: EISD860103
Summed mesophilic B protein values	Sum of B-values of mesophilic protein distributions of all residues	Extracted from AAindex [4]: PARS000101
Average mesophilic B protein values	Sum of B-values of mesophilic protein distributions divided by chain length	Extracted from AAindex [4]: PARS000101
Summed thermophilic B protein values	Sum of B-values of thermophilic protein distributions of all residues	Extracted from AAindex [4]: KUMS000101
Average thermophilic B protein values	Sum of B-values of thermophilic protein distributions divided by chain length	Extracted from AAindex [4]: KUMS000101
Summed buried fractions	Sum of ratio of buried and accessible molar fractions of all residues	Extracted from AAindex [4]: JANJ790101
Average buried fractions	Sum of ratio of buried and accessible molar fractions divided by chain length	Extracted from AAindex [4]: JANJ790101
Summed normalized flexibility	Sum of normalized flexibility parameters of all residues	Extracted from AAindex [4]: VINM940103
Average normalized flexibility	Sum of normalized flexibility parameters divided by chain length	Extracted from AAindex [4]: VINM940103
Total average normalized flexibility	Sum of averaged normalized flexibility parameters of all residues	Extracted from AAindex [4]: VINM940101
Average normalized flexibility	Sum of averaged normalized flexibility parameters divided by chain length	Extracted from AAindex [4]: VINM940101
Summed beta sheet frequency	Sum of normalized frequency of beta-sheet values of all residues	Extracted from AAindex [4]: PALJ810104
Average beta sheet frequency	Sum of normalized frequency of beta-sheet values divided by chain length	Extracted from AAindex [4]: PALJ810104
Summed 14A contact values	Sum of 14A contact numbers of all residues	Extracted from AAindex [4]: NISK860101
Average 14A contact values	Sum of 14A contact numbers divided by chain length	Extracted from AAindex [4]: NISK860101
Summed beta position 1 affinity	Sum of weights for beta-sheet at the window position of 1 of all residues.	Extracted from AAindex [4]: QIAN880121
Average beta position 1 affinity	Sum of weights for beta-sheet at the window position of 1 divided by chain length	Extracted from AAindex [4]: QIAN880121
Summed bilayer energy	Sum of free energies of transfer from bilayer interface to water values of all residues	Extracted from AAindex [4]: WIMW960101
Average bilayer energy	Sum of free energies of transfer from bilayer interface to water values divided by chain length	Extracted from AAindex [4]: WIMW960101
Summed normalized frequency of beta structures	Sum of normalized frequency of beta-structure values of all residues	Extracted from AAindex [4]: NAGK730102

Average normalized frequency of beta structures	Sum of normalized frequency of beta-structure values divided by chain length	Extracted from AAindex [4]: NAGK730102
Summed optimized side chains	Sum of side chain interaction parameter of all residues	Extracted from AAindex [4]: OOBM850105
Average optimized side chains	Sum of side chain interaction parameter divided by chain length	Extracted from AAindex [4]: OOBM850105
Summed occupancy of water	Sum of fraction of sites occupied by water of all residues	Extracted from AAindex [4]: KRIW790102
Average occupancy of water	Sum of fraction of site occupied by water divided by chain length	Extracted from AAindex [4]: KRIW790102
Summed normalized beta sheets	Sum of fraction of site normalized frequency of beta-sheets of all residues	Extracted from AAindex [4]: CHOP780202
Average normalized beta sheets	Sum of fraction of site normalized frequency of beta-sheets divided by chain length	Extracted from AAindex [4]: CHOP780202
Summed refractivity	Sum of refractivity of all residues	Extracted from AAindex [4]: MCMT640101
Average refractivity	Sum of refractivity divided by chain length	Extracted from AAindex [4]: MCMT640101
Summed bulkiness	Sum of bulkiness of all residues	Extracted from AAindex [4]: ZIMJ680102
Average bulkiness	Sum of bulkiness divided by chain length	Extracted from AAindex [4]: ZIMJ680102

Appendix 2 –DisoLipPred supplementary data

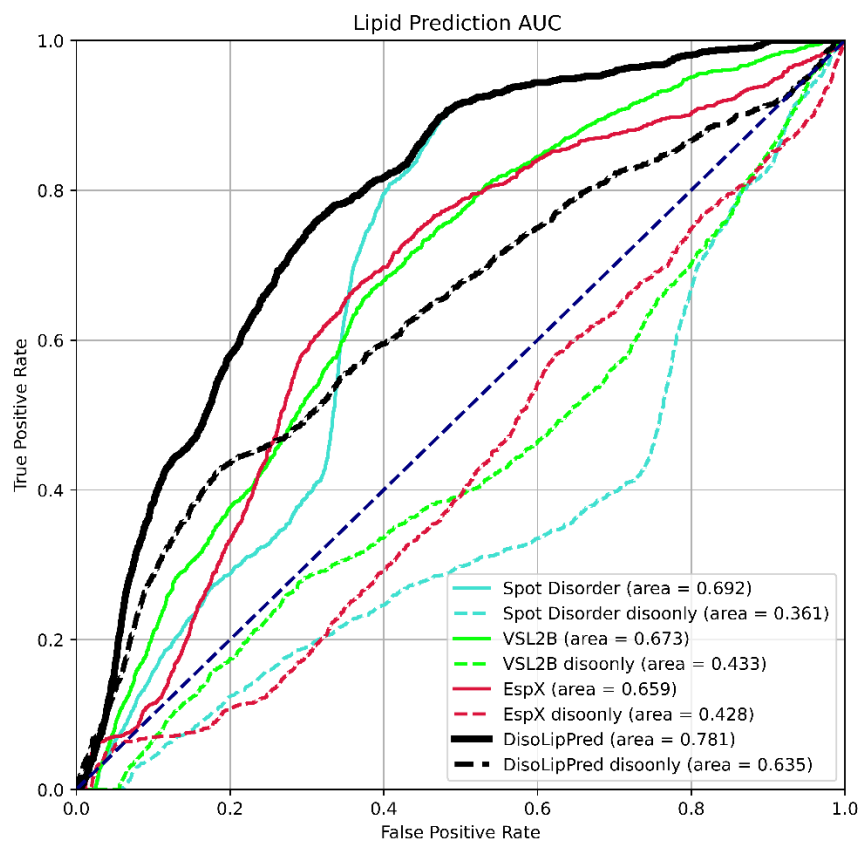
Appendix 2 Table 1: Partner-agnostic sequence profile

Description	Source
Predicted disorder propensity	Predicted with SPOT-Disorder [198]
Predicted solvent accessibility	Predicted with ASAquick [243]
Predicted coil propensity	Predicted with PSIPRED [147]
Predicted helix propensity	Predicted with PSIPRED [147]
Predicted strand propensity	Predicted with PSIPRED [147]
Predicted disordered protein binding propensity	Predicted with DisoRDPbind [116]
Predicted disordered DNA binding propensity	Predicted with DisoRDPbind [116]
Predicted disordered RNA binding propensity	Predicted with DisoRDPbind [116]
Predicted flexible linker propensity	Predicted with DFLpred [107]
Predicted disordered protein binding propensity	Predicted with ANCHOR [195]

Appendix 2 Table 2: Extended profile for the prediction of DLBRs.

Description	Source
Predicted disorder propensity	Predicted with SPOT-Disorder [198]
Predicted solvent accessibility	Predicted with ASAquick [243]
Predicted coil propensity	Predicted with PSIPRED [147]
Predicted helix propensity	Predicted with PSIPRED [147]
Predicted strand propensity	Predicted with PSIPRED [147]
Hydropathy	Extracted from AAindex [148]: KYTJ820101
Net charge	Extracted from AAindex [148]: KLEP840101
Polarity	Extracted from AAindex [148]: GRAR740102
Unfolding Gibbs energy values in water	Extracted from AAindex [148]: YUTK870101
Transfer energy	Extracted from AAindex [148]: OOBM850103
Solvation free energy	Extracted from AAindex [148]: EISD860101
Absolute entropy	Extracted from AAindex [148]: HUTJ700102
Isoelectric point	Extracted from AAindex [148]: ZIMJ680104
Charge transfer	Extracted from AAindex [148]: CHAM830107

Charge donor	Extracted from AAindex [148]: CHAM830108
Positive charge	Extracted from AAindex [148]: FAUJ880111
Negative charge	Extracted from AAindex [148]: FAUJ880112
Argos hydrophobicity	Extracted from AAindex [148]: ARGP820101
Kyte-Doolittle hydrophobicity	Extracted from AAindex [148]: JURD980101
Manavalan-Ponnuswamy hydrophobicity	Extracted from AAindex [148]: MANP780101
Cowan-Whittaker hydrophobicity	Extracted from AAindex [148]: COWR900101
Casari-Sippl hydrophobicity	Extracted from AAindex [148]: CASG920101
Alpha-CH chemical shifts	Extracted from AAindex [148]: ANDN920101
Spin-spin coupling constants	Extracted from AAindex [148]: GRAR740103
Membrane preference	Extracted from AAindex [148]: DESM900101
Atom-based hydrophobic moment	Extracted from AAindex [148]: EISD860102
Direction of the hydrophobic moment	Extracted from AAindex [148]: EISD860103
B-values	Extracted from AAindex [148]: PARS000101
Distribution frequencies in thermophilic proteins	Extracted from AAindex [148]: KUMS000101
B-values for residues with a rigid neighbor	Extracted from AAindex [148]: VINM940103
14 A contact number	Extracted from AAindex [148]: NISK860101
Free energies of transfer peptides from bilayer interface to water	Extracted from AAindex [148]: WIMW960101
Optimized side chain interaction parameter	Extracted from AAindex [148]: OOBM850105
Fraction of site occupied by water	Extracted from AAindex [148]: KRIW790102
Partition coefficient for ionic strength	Extracted from AAindex [148]: ZASB820101
Side chain hydropathy corrected for solvation	Extracted from AAindex [148]: ROSM880102
Affinity to bind transmembrane regions	Extracted from AAindex [148]: NAKH900112
Solvation free energy	Extracted from AAindex [148]: EISD860101
Activation Gibbs energy of unfolding at pH 9.0	Extracted from AAindex [148]: YUTK870104
Relative preference value at N2	Extracted from AAindex [148]: RICJ880105
STERIMOL length of the side chain	Extracted from AAindex [148]: FAUJ880104
Transfer free energy from chx to oct	Extracted from AAindex [148]: RADA880104
Propensity for N-terminal turn	Extracted from AAindex [148]: ROBB760109
Side chain torsion angle	Extracted from AAindex [148]: LEVM760104
Ratio of average and computed composition	Extracted from AAindex [148]: NAKH900113
Helix initiation parameter	Extracted from AAindex [148]: FINA910101
Pleated-sheet propensity	Extracted from AAindex [148]: ROBB760106
AA composition of mt-proteins from fungi and plant	Extracted from AAindex [148]: NAKH900107
Alpha-helix propensity	Extracted from AAindex [148]: KOEP990101
Alpha-helix propensity for alpha/beta-proteins	Extracted from AAindex [148]: GEIM800104
Normalized alpha-helix frequency	Extracted from AAindex [148]: MAXF760101



Supplementary Figure S1: ROC curves and AUC values on the test dataset for the prediction of DLBRs. Solid lines represent results on the complete test dataset while dashed lines show results on the native disordered residues in the test dataset.