Supplement for

"Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus"

Xiao Fan[1] and Lukasz Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

[*]corresponding author; lkurgan@ece.ualberta.ca; phone: 780-492-5488

Table S1. Summary of the considered disorder predictors. The constrained AUC is reported based on the predictions on the entire TRAINING dataset; except for MFDp for which the predictions are based on 5 fold cross validation. The methods are sorted by the constrained AUC in the descending order. The three types of availability are standalone program (SP), web server (WS) and upon request (UR).

Method	Constrained AUC	Year published	Type of disorder predictor	Availability	URL
ESpritz-Disprot	0.546	2012	machine learning	SP+WS	http://protein.bio.unipd.it/espritz/
CSpritz-long	0.540	2011	consensus	WS	http://protein.bio.unipd.it/cspritz/
SPINE-D	0.504	2012	machine learning	SP+WS	http://sparks.informatics.iupui.edu/SPINE-D/
CSpritz-short	0.499	2011	consensus	WS	http://protein.bio.unipd.it/cspritz/
MFDp	0.490	2010	consensus	SP+WS	http://biomine-ws.ece.ualberta.ca/MFDp.html
MD	0.476	2009	consensus	SP	https://rostlab.org/owiki/index.php/Metadisorder
IUPRED-short	0.465	2005	propensity	SP+WS	http://iupred.enzim.hu/
ESpritz-NMR	0.459	2012	machine learning	SP+WS	http://protein.bio.unipd.it/espritz/
IUPRED-long	0.458	2005	propensity	SP+WS	http://iupred.enzim.hu/
ESpritz-Xray	0.457	2012	machine learning	SP+WS	http://protein.bio.unipd.it/espritz/
DISOclust	0.456	2008	3D prediction	SP+WS	http://www.reading.ac.uk/bioinf/DISOclust/
VSL2B	0.453	2006	machine learning	SP+WS	http://www.dabi.temple.edu/disprot/Predictors.html
DISOPRED2	0.452	2003	machine learning	SP+WS	http://bioinf.cs.ucl.ac.uk/disopred/
PONDR-FIT	0.447	2010	consensus	UR	http://www.disprot.org/predictors.php
PrDos	0.439	2007	3D prediction	WS	http://prdos.hgc.jp/cgi-bin/top.cgi
RONN	0.420	2005	machine learning	SP+WS	http://www.strubi.ox.ac.uk/RONN
Norsnet	0.388	2007	machine learning	SP	https://www.rostlab.org/owiki/index.php/Norsnet
DRIP-PRED	0.377	2004	machine learning	SP+WS	https://www.sbc.su.se/~maccallr/disorder/
Ucon	0.367	2008	propensity	SP	https://rostlab.org/owiki/index.php/UCON
Profbval	0.334	2006	machine learning	SP	https://rostlab.org/owiki/index.php/Profbval

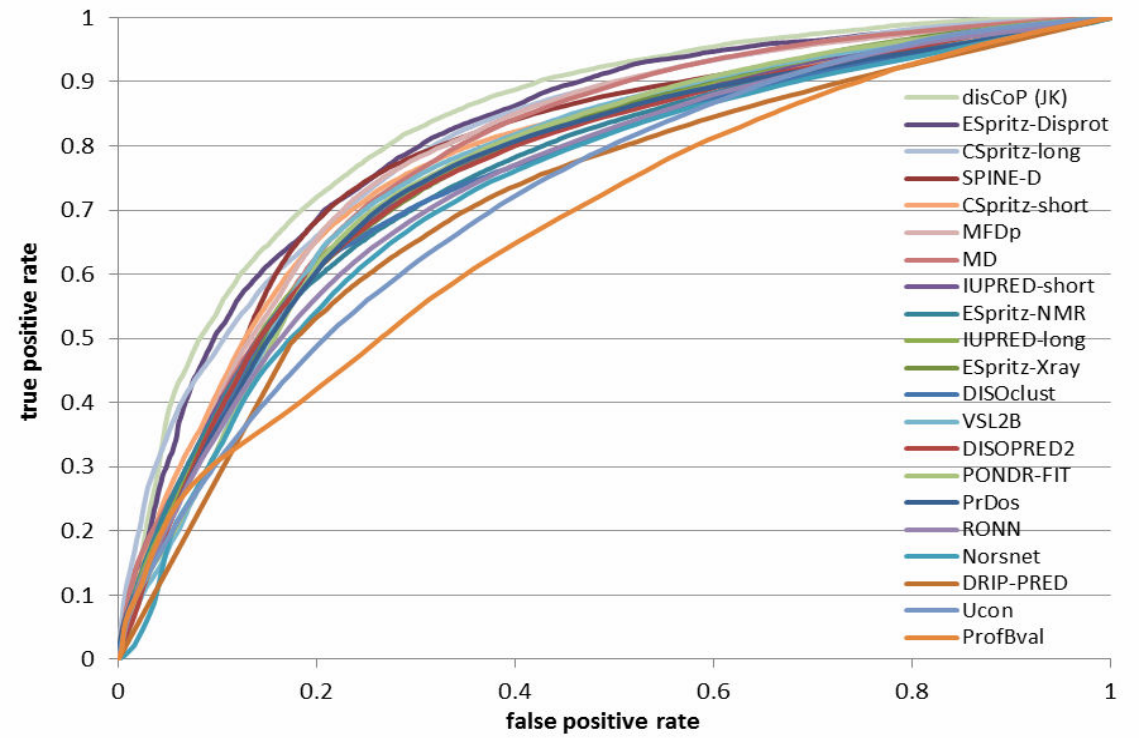

Figure S1. **ROC curves on the TRAINING dataset for the disCoP (based on jackknife test) and the other 20 predictors from the Supplementary Table S1.** The legend lists the methods that are sorted in the descending order by their values of the constrained AUC.

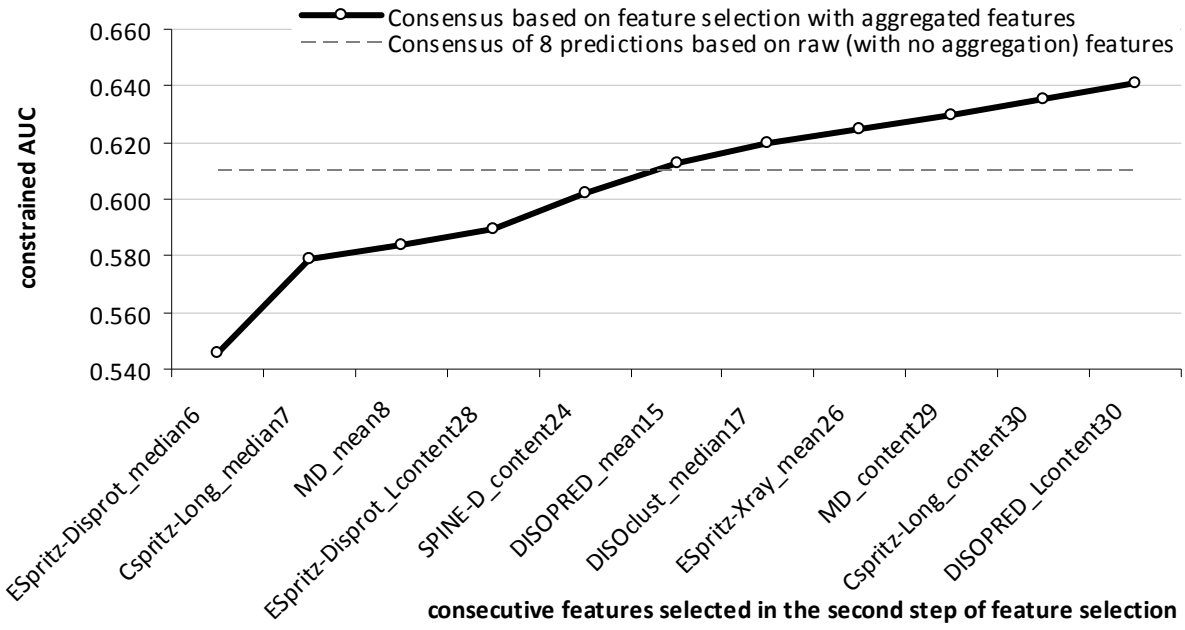

Figure S2. Values of the constrained AUCs with the increasing number of features selected in the second step of the feature selection. The results are based on three-fold cross validation on the TRAINING dataset. The first part of the feature name (*x*-axis) identifies the input predictor; the second part shows the particular type of output and aggregation where median*i* and mean*i* correspond to median and mean probability in window of size 2**i*+1, respectively, and content*i* and Lcontent*i* correspond to content of binary and ternary predictions in window of size 2**i*+1, respectively.

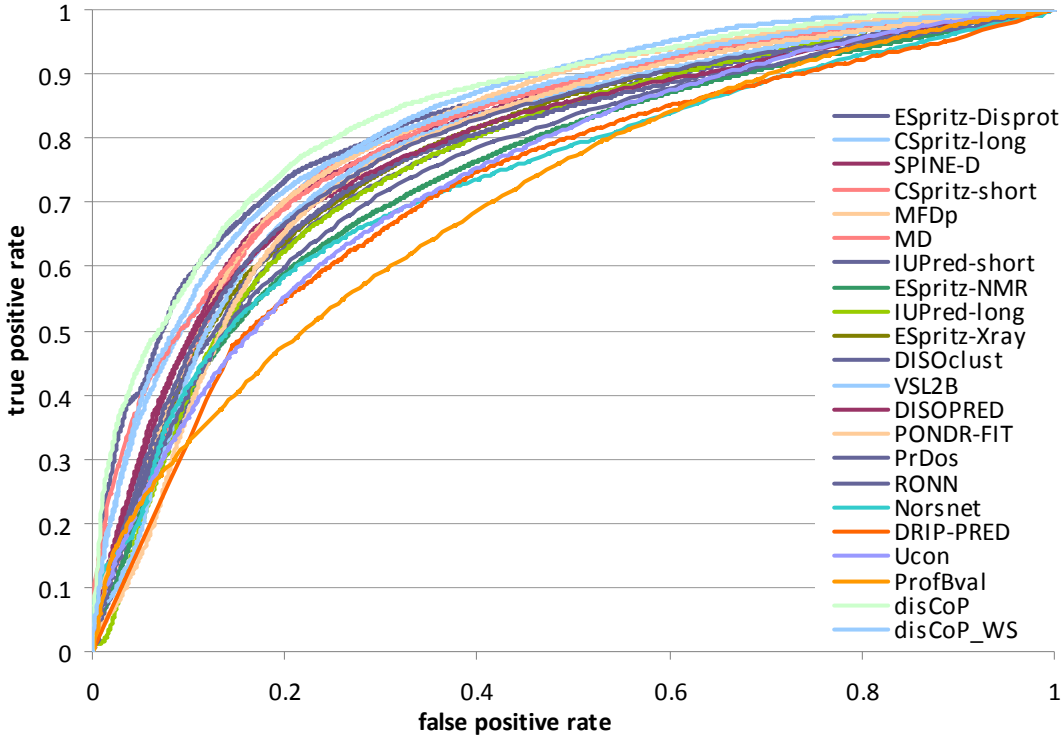

Figure S3. ROC curves of the disCoP, disCoP_WS and the other 20 predictors on the TEST dataset.

Table S2. Constrained AUC values measured on the TEST_FUNCTION dataset for disCoP, disCoP and 20 other predictors for the six functional types of disorder. The constrained AUC values are averages over the 10 repetitions with different randomly selected sets of structured residues (see "Datasets and evaluation protocols" section for details). Methods that are used as inputs to disCoP are shown in bold font. The highest value for each functional type is given in bold font. The methods are sorted by average rank of constrained AUC, which is the average over the ranks for individual functional types. disCoP_WS is a web server version of disCoP that excludes ESpritz and CSpritz predictors (see "disCoP predictor" section for details).

Method	Functional types related to binding			Other functional types			average constrained AUC	average rank of constrained AUC
	Protein-protein binding	Substrate or ligand binding	Protein-DNA binding	Flexible linkers or spacers	Phosphoryl-ation	Autoregula-tory		
ESpritz-Disprot	**0.689**	**0.504**	**0.574**	0.507	**0.733**	0.664	**0.612**	**1.33**
disCoP	0.642	0.481	0.492	**0.535**	0.709	**0.687**	0.591	1.83
disCoP server	0.583	0.415	0.486	0.502	0.697	0.631	0.553	2.83
MD	0.573	0.408	0.522	0.495	0.704	0.619	0.553	3.17
CSpritz-long	0.549	0.400	0.437	0.502	0.669	0.668	0.538	4.17
SPINE-D	0.529	0.395	0.412	0.490	0.651	0.621	0.516	6.17
CSpritz-short	0.471	0.397	0.425	0.485	0.569	0.569	0.486	8.17
MFDp	0.519	0.379	0.429	0.454	0.615	0.577	0.495	8.67
DISOPRED	0.494	0.377	0.384	0.475	0.623	0.612	0.494	8.83
ProDos	0.475	0.359	0.386	0.497	0.569	0.592	0.480	10.17
ESpritz-Xray	0.478	0.404	0.403	0.453	0.588	0.525	0.475	10.17
DISOCLUST	0.487	0.369	0.372	0.464	0.647	0.594	0.489	10.33
VSL2B	0.481	0.391	0.419	0.412	0.579	0.544	0.471	10.33
Norsnet	0.471	0.335	0.431	0.373	0.596	0.462	0.445	13.67
PONDR-FIT	0.455	0.374	0.370	0.409	0.514	0.463	0.431	14.83
RONN	0.434	0.351	0.366	0.408	0.529	0.488	0.429	15.17
Ucon	0.433	0.365	0.400	0.384	0.499	0.490	0.429	15.83
IUPRED-long	0.461	0.357	0.354	0.405	0.509	0.465	0.425	16.00
IUPRED-short	0.439	0.363	0.352	0.423	0.496	0.464	0.423	16.50
ESpritz-NMR	0.433	0.345	0.339	0.378	0.528	0.503	0.421	16.67
Profbval	0.381	0.343	0.370	0.351	0.470	0.450	0.394	18.50
DRIP-PRED	0.375	0.290	0.296	0.384	0.500	0.466	0.385	20.00

Table S3. Summary of 9 features used in the disCoP_WS consensus. The features are sorted by their constrained AUC when used individually to predict the disorder based on three-fold cross validation on the TRAINING dataset. The biserial correlation was computed against the native disorder annotation in the TRAINING dataset. The "Constrained AUC then added to consensus" gives the value of the constrained AUC when a given feature was added into the consensus during the feature selection. The last column lists weights in the regression including a bias (free weight), which is listed in the last row. The features with negative weights are given in bold font. The first part of the feature name (before underscore) identifies the input predictor; the second part shows the particular type of output and aggregation where median*i* and mean*i* correspond to median and mean probability in window of size 2**i*+1, respectively, and content*i* and Lcontent*i* correspond to content of binary and ternary predictions in window of size 2**i*+1, respectively (see "Feature generation and selection" for details).

Features	Predictive performance of individual features		Constrained AUC when added to consensus	Regression weights
	constrained AUC of individual features	Biserial correlation with native disorder		
SPINE-D_median12	0.508	0.460	0.508	0.186
MD_mean7	0.490	0.455	0.519	0.366
DISOCLUST_median13	0.472	0.390	0.526	0.050
SPINE-D_content14	0.472	0.451	0.532	0.192
DISOPRED_mean15	0.471	0.416	0.538	-0.036
DISOPRED_Lcontent22	0.437	0.386	0.546	-0.160
DISOCLUST_content29	0.425	0.367	0.554	-0.010
SPINE-D_median12	0.508	0.460	0.508	0.186
MD_mean7	0.490	0.455	0.519	0.366
bias				0.206