Supplementary Information for
"DISOselect: disorder predictor selection at the protein level"

Akila Katuwawala[1], Christopher Oldfield[1], and Lukasz Kurgan[1]*
[1]Department of Computer Science, Virginia Commonwealth University

*corresponding author:
Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, Virginia 23284, USA
Email: lkurgan@vcu.edu; Phone: (804) 827-3986

Supplementary Table S1. Description of the 130 features that were considered in the design of DISOselect.

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 (2)
	Average accessible surface area	Average of solvent accessibility of all residues	Predicted with ASAquick (2)
	Total number of exposed residues	Sum of binary exposed residues	Predicted with ASAquick (2)

Sequence Complexity (2 features)	Fraction of complex regions	Number of complex regions divided by chain length	Computed by SEG (4)
	Fraction of complex residues	Number of complex residues divided by chain length	Computed by SEG (4)
Predicted Secondary Structure (8 features)	Count of coils	Count of putative coil residues in protein	Predicted with PSIPRED (1)
	Count of helices	Count of putative helix residues in protein	Predicted with PSIPRED (1)
	Count of strands	Count of putative strand residues in protein	Predicted with PSIPRED (1)
	Count of coils and strands	Count of putative coil and strand residues in protein	Predicted with PSIPRED (1)
	Content of coils	Fraction of putative coil residues in the input protein chain	Predicted with PSIPRED (1)
	Content of helices	Fraction of putative helix residues in the input protein chain	Predicted with PSIPRED (1)
	Content of strands	Fraction of putative strands residues in the input protein chain	Predicted with PSIPRED (1)
	Content of coils and strands	Fraction of putative coils and strand residues in the input protein chain	Predicted with PSIPRED (1)
Physiochemical properties of amino acids (97 features)	Summed hydropathy	Sum of hydropathy values of all residues	Extracted from AAindex (3): KYTJ820101
	Summed net charge	Sum of net charge values of all residues	Extracted from AAindex (3): KLEP840101
	Summed hydrophilicity	Sum of hydrophilicity values of all residues of all residues	Extracted from AAindex (3): HOPT810101
	Average hydrophilicity	Sum of hydrophilicity values divided by chain length	Extracted from AAindex (3): HOPT810101
	Average absolute entropy	Sum of absolute entropy values divided by chain length	Extracted from AAindex (3): HUTJ700102
	Average unfolding gibbs energy	Sum of unfolding Gibbs energy values divided by chain length	Extracted from AAindex (3): YUTK870101
	Average beta coils	Sum of beta-structure-coil equilibrium constants divided by chain length	Extracted from AAindex (3): OOBM850101
	Average reverse turns	Sum of propensities to form reverse turn divided by chain length	Extracted from AAindex (3): OOBM850102
	Summed transfer energy	Sum of transfer energy parameters of all residues	Extracted from AAindex (3): OOBM850103
	Average Isoelectricity	Sum of isoelectric points divided by chain length	Extracted from AAindex (3): 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 (3): PRAM900101
	Average hydrophobicity	Sum of hydrophobicity values divided by chain length	Extracted from AAindex (3): PRAM900101
	Average hydropathy	Sum of hydropathy values divided by chain length	Extracted from AAindex (3): KYTJ820101
	Summed solvation free energy	Sum of solvation free energy values of all residues	Extracted from AAindex (3): EISD860101
	Average solvation free energy	Sum of solvation free energy values divided by chain length	Extracted from AAindex (3): EISD860101
	Summed polarity	Sum of polarity values of all residues	Extracted from AAindex (3): GRAR740102
	Average polarity	Sum of polarity values divided by chain length	Extracted from AAindex (3): GRAR740102
	Summed volume	Sum of volume values of all residues	Extracted from AAindex (3): GRAR740103
	Average volume	Sum of volume values divided by chain length	Extracted from AAindex (3): GRAR740103
	Summed absolute entropy	Sum of absolute entropy values of all residues	Extracted from AAindex (3): HUTJ700102

	Summed unfolding gibbs	Sum of unfolding Gibbs energy in water of all residues	Extracted from AAindex (3): YUTK870101
	Summed activation gibbs	Sum of activation Gibbs energy values of all residues	Extracted from AAindex (3): KLEP840101
	Average activation gibbs	Sum of activation Gibbs energy values divided by chain length	Extracted from AAindex (3): KLEP840101
	Summed beta coils	Sum of beta-structure-coil equilibrium constants of all residues	Extracted from AAindex (3): OOBM850101
	Summed reverse turn	Sum of propensity to form reverse turn values of all residues	Extracted from AAindex (3): OOBM850102
	Average transfer energy	Sum of transfer energy parameters divided by chain length	Extracted from AAindex (3): OOBM850103
	Summed isoelectric points	Sum of isoelectric points values of all residues	Extracted from AAindex (3): ZIMJ680104
	Summed charge transfer	Sum of parameter of charge transfer capability of all residues	Extracted from AAindex (3): CHAM830107
	Summed charge donor	Sum of parameter of charge donor capability of all residues	Extracted from AAindex (3): CHAM830108
	Summed positive charge	Sum of parameter of positive charge capability of all residues	Extracted from AAindex (3): CHAM830108
	Summed negative charge	Sum of parameter of negative charge capability of all residues	Extracted from AAindex (3): CHAM830108
	Summed hydrophobicity index	Sum of hydrophobicity Index values of all residues	Extracted from AAindex (3): ARGP820101
	Average hydrophobicity index	Sum of hydrophobicity Index values divided by chain length	Extracted from AAindex (3): ARGP820101
	Summed alpha hydrophobicity	Sum of normalized hydrophobicity scales for alpha-proteins of all residues	Extracted from AAindex (3): CIDH920101
	Average alpha hydrophobicity	Sum of normalized hydrophobicity scales for alpha-proteins divided by chain length	Extracted from AAindex (3): CIDH920101
	Summed normalized average hydrophobicity	Sum of normalized average hydrophobicity scales of all residues	Extracted from AAindex (3): CIDH920105
	Average normalized average hydrophobicity	Sum of Normalized average hydrophobicity scales divided by chain length	Extracted from AAindex (3): CIDH920105
	Summed consensus normalized hydrophobicity	Sum of consensus normalized hydrophobicity scales of all residues	Extracted from AAindex (3): EISD840101
	Average consensus normalized hydrophobicity	Sum of Consensus normalized hydrophobicity scales divided by chain length	Extracted from AAindex (3): EISD840101
	Summed average surrounding hydrophobicity	Sum of average surrounding hydrophobicity values of all residues	Extracted from AAindex (3): MANP780101
	Average surrounding hydrophobicity	Sum of average surrounding hydrophobicity values divided by chain length	Extracted from AAindex (3): MANP780101
	Summed hydrophobicityPH3	Sum of hydrophobicity index values at 3.0 pH of all residues	Extracted from AAindex (3): COWR900101

	Average hydrophobicityPH3	Sum of hydrophobicity index values at 3.0 pH divided by chain length	Extracted from AAindex (3): COWR900101
	Summed native hydrophobicity	Sum of native hydrophobicity index values of all residues	Extracted from AAindex (3): CASG920101
	Average native hydrophobicity	Sum of native Hydrophobicity index values divided by chain length	Extracted from AAindex (3): 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 (3): ANDN920101
	Average CH chemical shifts	Sum of alphaCH chemical shift values divided by chain length	Extracted from AAindex (3): ANDN920101
	Summed NH chemical shifts	Sum of alphaNH chemical shift values of all residues	Extracted from AAindex (3): BUNA790101
	Average NH chemical shifts	Sum of alphaNH chemical shift values divided by chain length	Extracted from AAindex (3): BUNA790101
	Summed spin coupling	Sum of spin coupling constants of all residues	Extracted from AAindex (3): BUNA790103
	Average spin coupling	Sum of spin coupling constants divided by chain length	Extracted from AAindex (3): BUNA790103
	Summed membrane preference	Sum of membrane preference indexes of all residues	Extracted from AAindex (3): DESM900101
	Average membrane preference	Sum of membrane preference indexes divided by chain length	Extracted from AAindex (3): DESM900101
	Summed hydrophobic moment	Sum of atom-based hydrophobic moment values of all residues	Extracted from AAindex (3): EISD860102
	Average hydrophobic moment	Sum of atom-based hydrophobic moment values divided by chain length	Extracted from AAindex (3): EISD860102
	Summed hydrophobic moment direction	Sum of direction of hydrophobic moment values of all residues	Extracted from AAindex (3): EISD860103
	Average hydrophobic moment direction	Sum of direction of hydrophobic moment values divided by chain length	Extracted from AAindex (3): EISD860103

	Summed mesophilic B protein values	Sum of B-values of mesophilic protein distributions of all residues	Extracted from AAindex (3): PARS000101
	Average mesophilic B protein values	Sum of B-values of mesophilic protein distributions divided by chain length	Extracted from AAindex (3): PARS000101
	Summed thermophilic B protein values	Sum of B-values of thermophilic protein distributions of all residues	Extracted from AAindex (3): KUMS000101
	Average thermophilic B protein values	Sum of B-values of thermophilic protein distributions divided by chain length	Extracted from AAindex (3): KUMS000101
	Summed buried fractions	Sum of ratio of buried and accessible molar fractions of all residues	Extracted from AAindex (3): JANJ790101
	Average buried fractions	Sum of ratio of buried and accessible molar fractions divided by chain length	Extracted from AAindex (3): JANJ790101
	Summed normalized flexibility	Sum of normalized flexibility parameters of all residues	Extracted from AAindex (3): VINM940103
	Average normalized flexibility	Sum of normalized flexibility parameters divided by chain length	Extracted from AAindex (3): VINM940103
	Total average normalized flexibility	Sum of averaged normalized flexibility parameters of all residues	Extracted from AAindex (3): VINM940101
	Average normalized flexibility	Sum of averaged normalized flexibility parameters divided by chain length	Extracted from AAindex (3): VINM940101
	Summed beta sheet frequency	Sum of normalized frequency of beta-sheet values of all residues	Extracted from AAindex (3): PALJ810104
	Average beta sheet frequency	Sum of normalized frequency of beta-sheet values divided by chain length	Extracted from AAindex (3): PALJ810104
	Summed 14A contact values	Sum of 14A contact numbers of all residues	Extracted from AAindex (3): NISK860101
	Average 14A contact values	Sum of 14A contact numbers divided by chain length	Extracted from AAindex (3): NISK860101
	Summed beta position 1 affinity	Sum of weights for beta-sheet at the window position of 1 of all residues.	Extracted from AAindex (3): 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 (3): QIAN880121
	Summed bilayer energy	Sum of free energies of transfer from bilayer interface to water values of all residues	Extracted from AAindex (3): WIMW960101
	Average bilayer energy	Sum of free energies of transfer from bilayer interface to water values divided by chain length	Extracted from AAindex (3): WIMW960101
	Summed normalized frequency of beta structures	Sum of normalized frequency of beta-structure values of all residues	Extracted from AAindex (3): NAGK730102
	Average normalized frequency of beta structures	Sum of normalized frequency of beta-structure values divided by chain length	Extracted from AAindex (3): NAGK730102
	Summed optimized side chains	Sum of side chain interaction parameter of all residues	Extracted from AAindex (3): OOBM850105

	Average optimized side chains	Sum of side chain interaction parameter divided by chain length	Extracted from AAindex (3): OOBM850105
	Summed occupancy of water	Sum of fraction of sites occupied by water of all residues	Extracted from AAindex (3): KRIW790102
	Average occupancy of water	Sum of sraction of site occupied by water divided by chain length	Extracted from AAindex (3): KRIW790102
	Summed normalized beta sheets	Sum of fraction of site normalized frequency of beta-sheets of all residues	Extracted from AAindex (3): CHOP780202
	Average normalized beta sheets	Sum of fraction of site normalized frequency of beta-sheets divided by chain length	Extracted from AAindex (3): CHOP780202
	Summed refractivity	Sum of refractivity of all residues	Extracted from AAindex (3): MCMT640101
	Average refractivity	Sum of refractivity divided by chain length	Extracted from AAindex (3): MCMT640101
	Summed bulkiness	Sum of bulkiness of all residues	Extracted from AAindex (3): ZIMJ680102
	Average bulkiness	Sum of bulkiness divided by chain length	Extracted from AAindex (3): ZIMJ680102

Supplementary Table S2. Predictive performance of the considered three types of machine learning models algorithms reported based on the 3-fold cross validation experiments on the training dataset. The performance is measured with Pearson correlation coefficient (PCC), mean absolute error (MAE) and mean squared error (MSE) between the predicted and the actual AUC values. The last line shows the average values across the 12 considered disorder predictors.

Disorder predictor	**Nearest Neighbor regressor**			**Linear regressor**			**Extra-tree regressor**		
	PCC	**MAE**	**MSE**	**PCC**	**MAE**	**MSE**	**PCC**	**MAE**	**MSE**
disEMBL-465	0.211	0.125	0.018	0.160	0.120	0.010	0.315	0.101	0.007
disEMBL-HL	0.203	0.141	0.021	0.244	0.135	0.011	0.384	0.123	0.008
ESpritz-DisProt	0.132	0.192	0.016	0.301	0.181	0.013	0.411	0.133	0.011
ESpritz-NMR	0.129	0.136	0.019	0.152	0.126	0.009	0.362	0.114	0.006
ESpritz-Xray	0.151	0.124	0.017	0.141	0.111	0.008	0.296	0.110	0.006
GlobPlot	0.123	0.147	0.019	0.293	0.136	0.011	0.304	0.125	0.008
IUPred-long	0.176	0.202	0.018	0.193	0.183	0.012	0.271	0.169	0.007
IUPred-short	0.242	0.147	0.020	0.252	0.145	0.013	0.297	0.143	0.008
JRONN	-0.056	0.209	0.018	0.127	0.189	0.011	0.207	0.177	0.005
VSL2B	0.187	0.129	0.017	0.180	0.127	0.010	0.328	0.113	0.006
SPOT-Disorder	0.124	0.072	0.018	0.107	0.068	0.009	0.254	0.066	0.006
DISOPRED3	0.191	0.078	0.011	0.120	0.074	0.014	0.261	0.063	0.006
Average of the 12 disorder predictors	**0.151**	**0.142**	**0.018**	**0.189**	**0.133**	**0.011**	**0.308**	**0.120**	**0.007**

Supplementary Table S3. Improvements in the dataset-level AUC values based on the selection of proteins using the putative AUC values generated by DISOselect for each of the 12 considered disorder predictors. The disorder predictors are sorted by their complete test dataset-level AUCs. For each disorder predictor the results on the whole test set are compared against the 75%, 50% and 25% proteins that have the highest putative AUC values. Significance of the differences between the AUCs on the whole test dataset and the AUCs for each of the three subsets was evaluated 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 each dataset ten times at random and compared the corresponding 10 pairs of AUCs; the resulting *p*-values are reported in the table inside the square brackets.

Disorder predictor	Datasets			
	Complete	Top 75% selected with DISOselect [*p*-value]	Top 50% selected with DISOselect [*p*-value]	Top 25% selected with DISOselect [*p*-value]
SPOT-Disorder	0.918	0.942 [<.01]	0.944 [<.01]	0.951 [<.01]
DISOPRED3	0.915	0.941 [<.01]	0.952 [<.01]	0.958 [<.01]
VSL2B	0.839	0.861 [<.01]	0.864 [<.01]	0.893 [<.01]
ESpritz-Xray	0.812	0.844 [<.01]	0.855 [<.01]	0.867 [<.01]
IUPred-short	0.810	0.836 [<.01]	0.845 [<.01]	0.856 [<.01]
ESpritz-NMR	0.807	0.824 [<.01]	0.852 [<.01]	0.878 [<.01]
disEMBL-465	0.804	0.832 [<.01]	0.844 [<.01]	0.875 [<.01]
ESpritz-DisProt	0.782	0.805 [<.01]	0.821 [<.01]	0.842 [<.01]
JRONN	0.766	0.776 [<.01]	0.789 [<.01]	0.817 [<.01]
disEMBL-HL	0.761	0.799 [<.01]	0.820 [<.01]	0.842 [<.01]
IUPred-long	0.732	0.762 [<.01]	0.777 [<.01]	0.777 [<.01]
GlobPlot	0.630	0.649 [<.01]	0.655 [<.01]	0.667 [<.01]

References

1. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Research 41:W349-W357.
2. Faraggi E, Zhou Y, Kloczkowski A (2014) Accurate single-sequence prediction of solvent accessible surface area using local and global features. Proteins 82:3170-3176. PMID: 25204636 {Medline}
3. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202-205. PMID: 17998252 {Medline}
4. Wootton JC (1994) Non-globular domains in protein sequences: Automated segmentation using complexity measures. Computers & Chemistry 18:269-285.