Supplement for "DisoLipPred: Accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning"

Akila Katuwawala[1], Bi Zhao[1], Lukasz Kurgan[1*]
[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

Supplementary Tables

Supplementary Table S1. Description of the training and test datasets.

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

Supplementary Table S2. Partner-agnostic sequence profile.

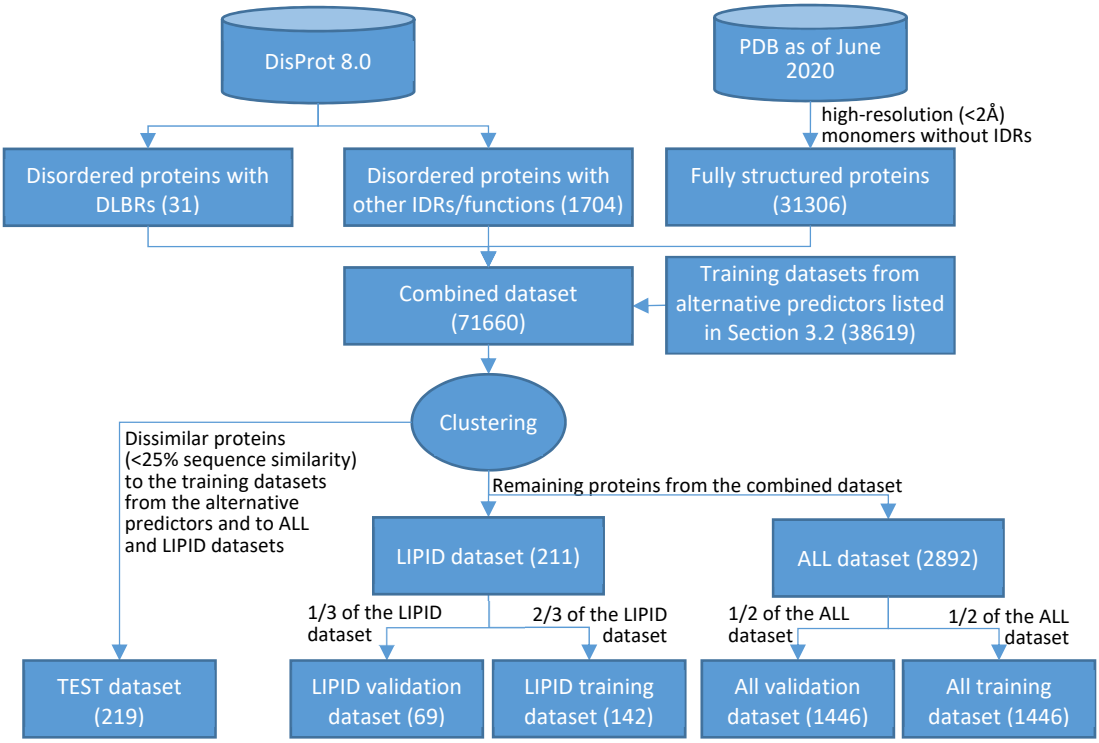
Description	Source
Predicted disorder propensity	Predicted with SPOT-Disorder [1]
Predicted solvent accessibility	Predicted with ASAquick [2]
Predicted coil propensity	Predicted with PSIPRED [3]
Predicted helix propensity	Predicted with PSIPRED [3]
Predicted strand propensity	Predicted with PSIPRED [3]
Predicted disordered protein binding propensity	Predicted with DisoRDPbind [4]
Predicted disordered DNA binding propensity	Predicted with DisoRDPbind [4]
Predicted disordered RNA binding propensity	Predicted with DisoRDPbind [4]
Predicted flexible linker propensity	Predicted with DFLpred [5]
Predicted disordered protein binding propensity	Predicted with ANCHOR [6]

Supplementary Table S3. Extended profile for the prediction of DLBRs.

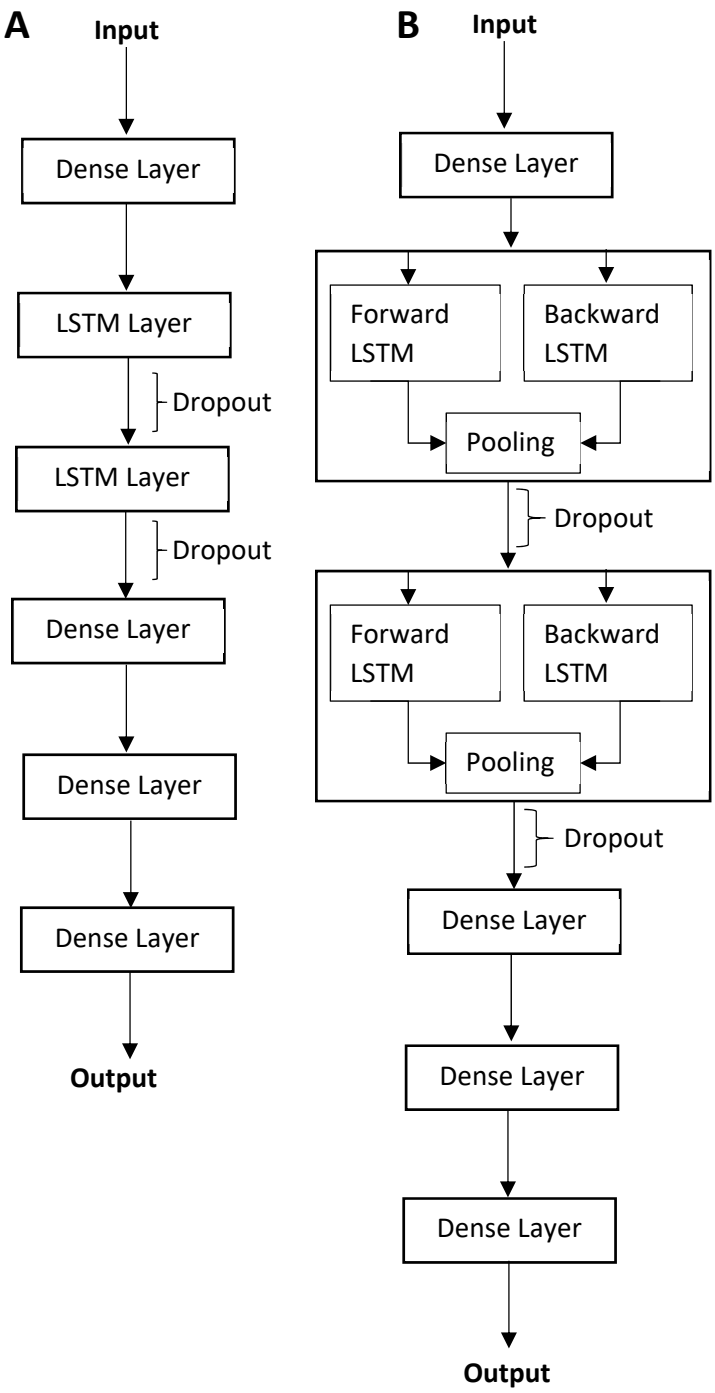
Description	Source
Predicted disorder propensity	Predicted with SPOT-Disorder [1]
Predicted solvent accessibility	Predicted with ASAquick [2]
Predicted coil propensity	Predicted with PSIPRED [3]
Predicted helix propensity	Predicted with PSIPRED [3]
Predicted strand propensity	Predicted with PSIPRED [3]
Hydropathy	Extracted from AAindex [9]: KYTJ820101
Net charge	Extracted from AAindex [9]: KLEP840101
Polarity	Extracted from AAindex [9]: GRAR740102
Unfolding Gibbs energy values in water	Extracted from AAindex [9]: YUTK870101
Transfer energy	Extracted from AAindex [9]: OOBM850103
Solvation free energy	Extracted from AAindex [9]: EISD860101
Absolute entropy	Extracted from AAindex [9]: HUTJ700102
Isoelectric point	Extracted from AAindex [9]: ZIMJ680104
Charge transfer	Extracted from AAindex [9]: CHAM830107
Charge donor	Extracted from AAindex [9]: CHAM830108
Positive charge	Extracted from AAindex [9]: FAUJ880111
Negative charge	Extracted from AAindex [9]: FAUJ880112
Argos hydrophobicity	Extracted from AAindex [9]: ARGP820101
Kyte-Doolittle hydrophobicity	Extracted from AAindex [9]: JURD980101
Manavalan-Ponnuswamy hydrophobicity	Extracted from AAindex [9]: MANP780101
Cowan-Whittaker hydrophobicity	Extracted from AAindex [9]: COWR900101
Casari-Sippl hydrophobicity	Extracted from AAindex [9]: CASG920101
Alpha-CH chemical shifts	Extracted from AAindex [9]: ANDN920101
Spin-spin coupling constants	Extracted from AAindex [9]: GRAR740103
Membrane preference	Extracted from AAindex [9]: DESM900101
Atom-based hydrophobic moment	Extracted from AAindex [9]: EISD860102
Direction of the hydrophobic moment	Extracted from AAindex [9]: EISD860103
B-values	Extracted from AAindex [9]: PARS000101
Distribution frequencies in thermophilic proteins	Extracted from AAindex [9]: KUMS000101
B-values for residues with a rigid neighbor	Extracted from AAindex [9]: VINM940103
14 A contact number	Extracted from AAindex [9]: NISK860101
Free energies of transfer peptides from bilayer interface to water	Extracted from AAindex [9]: WIMW960101
Optimized side chain interaction parameter	Extracted from AAindex [9]: OOBM850105
Fraction of site occupied by water	Extracted from AAindex [9]: KRIW790102
Partition coefficient for ionic strength	Extracted from AAindex [9]: ZASB820101
Side chain hydropathy corrected for solvation	Extracted from AAindex [9]: ROSM880102
Affinity to bind transmembrane regions	Extracted from AAindex [9]: NAKH900112
Solvation free energy	Extracted from AAindex [9]: EISD860101
Activation Gibbs energy of unfolding at pH 9.0	Extracted from AAindex [9]: YUTK870104
Relative preference value at N2	Extracted from AAindex [9]: RICJ880105
STERIMOL length of the side chain	Extracted from AAindex [9]: FAUJ880104
Transfer free energy from chx to oct	Extracted from AAindex [9]: RADA880104
Propensity for N-terminal turn	Extracted from AAindex [9]: ROBB760109
Side chain torsion angle	Extracted from AAindex [9]: LEVM760104
Ratio of average and computed composition	Extracted from AAindex [9]: NAKH900113
Helix initiation parameter	Extracted from AAindex [9]: FINA910101
Pleated-sheet propensity	Extracted from AAindex [9]: ROBB760106
AA composition of mt-proteins from fungi and plant	Extracted from AAindex [9]: NAKH900107
Alpha-helix propensity	Extracted from AAindex [9]: KOEP990101
Alpha-helix propensity for alpha/beta-proteins	Extracted from AAindex [9]: GEIM800104
Normalized alpha-helix frequency	Extracted from AAindex [9]: MAXF760101

Supplementary Table S4. Predictive performance of DisoLipPred and its variants from the ablation analysis (Table 1) on the validation dataset. We perform the assessment on the complete validation dataset, and also on the subset of disordered residues from the validation dataset. We quantify the binary metrics (sensitivity and F1) at the fixed specificity = 0.9. This enables direct comparison of the binary metrics between different variants.

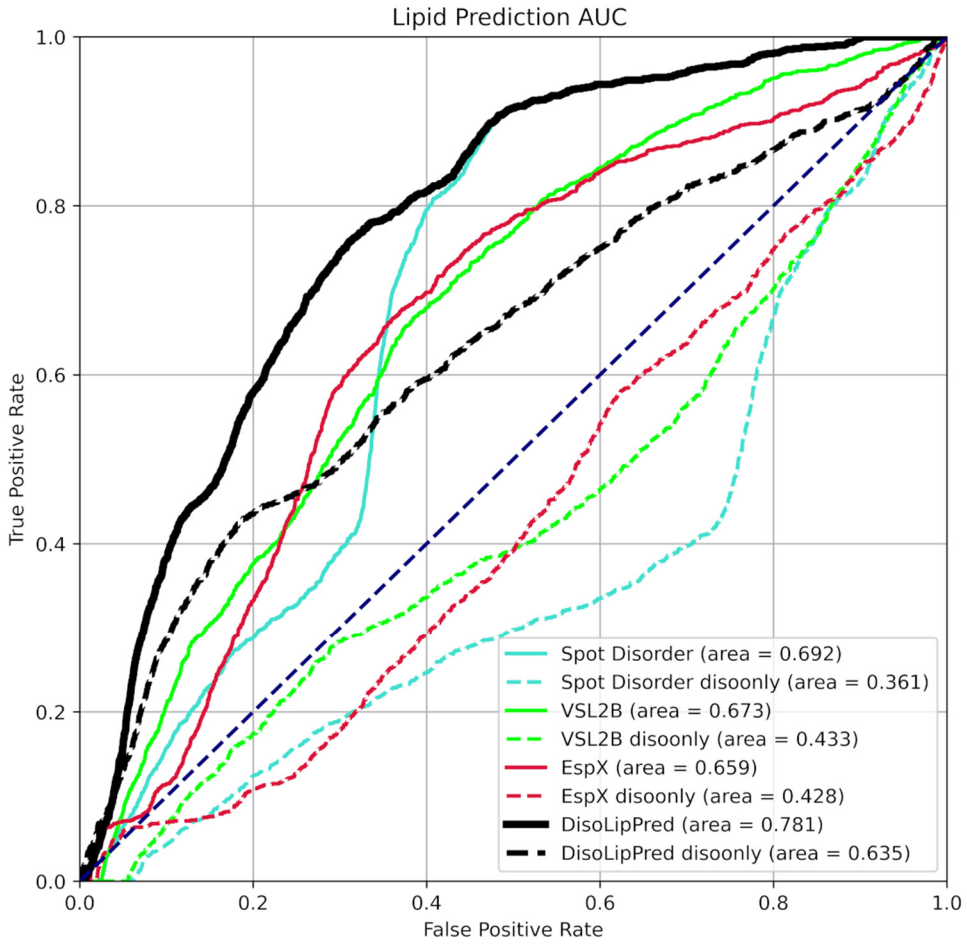
	Complete validation dataset				Disordered residues in the validation dataset			
Setup	**AUC**	**Sensitivity**	**Specificity**	**F1**	**AUC**	**Sensitivity**	**Specificity**	**F1**
DisoLipPred	0.809	0.228	0.900	0.149	0.708	0.351	0.900	0.190
1	0.769	0.221	0.900	0.081	0.701	0.313	0.900	0.172
2	0.710	0.080	0.900	0.072	0.660	0.128	0.900	0.074
3	0.669	0.206	0.900	0.077	0.564	0.241	0.900	0.134
4	0.600	0.120	0.900	0.039	0.500	0.088	0.900	0.052



Supplementary Figure S1: Flowchart for the generation of training, validation and test datasets. The numbers in the round brackets give the numbers of proteins in a given protein set.



Supplementary Figure S2. 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.



Supplementary Figure S3. 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.

References

1. 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.
2. 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.
3. Buchan, D.W.A., et al., *Scalable web services for the PSIPRED Protein Analysis Workbench.* Nucleic Acids Research, 2013. **41**(W1): p. W349-W357.
4. 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.
5. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences.* Bioinformatics, 2016. **32**(12): p. i341-i350.
6. 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.
7. Hanson, J., K. Paliwal, and Y. Zhou, *Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures.* Journal of Chemical Information and Modeling, 2018. **58**(11): p. 2369-2376.
8. 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.
9. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic acids research, 2008. **36**(Database issue): p. D202-D205.