Supplement for

SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences

Jian Zhang[1,2], and Lukasz Kurgan[2,*]

[1]School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000, [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, 23284.

*To whom correspondence should be addressed. Phone: +1-804-827-3986; Fax: +1-804-828-2771; Email: lkurgan@vcu.edu

Supplementary Table S1. Comparison of sizes of training and test datasets used to develop SCRIBER and the other seven sequence-based predictors of PBRs. Datasets used by the SPRINT method were undersampled at the residue level with the goal of using equal number of PBRs and non-binding residues, resulting in a relatively small set of residues used.

Method	Training dataset		Testing dataset	
	Number of proteins	Number of residues	Number of proteins	Number of residues
SPPIDER	435	98,285	149	28,021
PSIVER	186	36,219	72	18,140
LORIS	186	36,219	236	51,821
SPRINGS	186	36,219	236	51,821
CRF-PPI	186	36,219	236	51,821
SSWRF	186	36,219	236	51,821
SPRINT	1199	31,376	80	2,102
SCRIBER	843	225,299	448	116,500

Supplementary Table S2. Description of features that are utilized in the first layer of the SCRIBER. The features are computed in two ways: 1) for individual amino acids when using window size = 5 (i.e., values for each of the five residues in the window); and 2) based on values aggregated over the entire window for window size = 11 (typically average and standard deviations for the 11 values within the window).

Feature group	Feature type	Description	Window Size	Number of features	Number of features per feature type
Features derived from previously used inputs	Relative Amino Acid Propensity (RAAP) for binding	RAAP for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding versus non-binding (per residue)	5	5×7=35	168
		RAAP for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding versus other-binding (per residue)	5	5×7=35	
		RAAP for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding versus not {protein, DNA, RNA, small ligands, RNA+DNA+small ligands}-binding (per residue)	5	5×7=35	
		Fraction of residues with high (above average) propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within the window	11	3×7=21	
		Average and standard deviation of RAAP values for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within the window	11	2×3×7=42	
	Putative Relative Solvent Accessibility (RSA)	Putative RSA values (per residue)	5	5×1=5	9
		Fraction of putative exposed residues within the window (solvent exposed defined as RSA > 5% and as RSA > 20%)	11	2	
		Average and standard deviation of putative RSA values within the window	11	2	

	Evolutionary Conservation (ECO)	ECO scores (per residue)	5	5×1=5	9
		Fraction of conserved residues within the window (conserved is defined as ECO > average in TRAINING dataset and as ECO > average + stdev in TRAINING dataset)	11	2	
		Average and standard deviation of ECO scores within the window	11	2	
	Hydrophobicity, Polarity & Charge	Hydrophobicity values using two amino acid hydrophobicity indices (per residue)	5	5×2=10	46
		Average and standard deviation of hydrophobicity values within the window using two amino acid hydrophobicity indices	11	2×2=4	
		Fraction of hydrophobic residues within the window	11	1	
		Polarity values using two polarity indices (per residue)	5	5×2=10	
		Average and standard deviation of polarity values within the window using two polarity indices	11	2×2=4	
		Fraction of polar residues within the window	11	1	
		Positive and negative charge values (per residue)	5	5×2=10	
		Average and standard deviation of positive and negative charge values within the window	11	2×2=4	
		Fraction of positively and negatively charged residues within the window	11	2	
Features derived from novel inputs	Putative protein-binding disordered regions	Putative protein-binding disorder probability (per residue)	5	5×1=5	10
		Fraction of putative protein-binding disordered residues within the window	11	1	
		Average and standard deviation of putative protein-binding disorder probabilities within the window	11	1	
		Fraction of residues in the longest putative protein-binding disordered region within the window	11	1	
		Fraction of residues in the longest putative non-protein-binding disordered region within the window	11	1	
		Fraction of residues in the longest putative protein-binding disordered and non-protein-binding regions within that window	11	1	
	Putative Secondary Structure	Putative secondary structure coded using three bits (for helix, beta-sheet and coil) (per residue)	5	5×3=15	65
		Fraction of residues in putative helix conformation, in putative coil conformation, and in putative beta-sheet conformation within the window	11	3	
		Fraction of residues in putative helix and beta sheet conformations within the window	11	1	
		Fraction of residues in the longest putative helix segment, the longest putative beta sheet segment, and the longest putative coil segment within the window	11	3	
		Residue position within current putative secondary segment (linear distance from the terminus of the current secondary structure segment)	N/A	1	
		Average length of putative secondary structure segments in the sequence	N/A	1	
		Fraction of residues in putative helix conformation, in putative coil conformation, and in putative beta-sheet conformation in the whole sequence	N/A	3	
		Presence of a secondary structure motif at the position of the predicted residue coded using fourteen bits per residue (X[H|E|C](the head motif), [H|E|C]X(the tail motif), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, and HEH)	N/A	14	
		Fraction of residues in a given motif type in the sequence (XHE,XHC,XEH,XEC, XCH, XCE, HEX, HCX, EHX, ECX, EHX, ECX, CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH)	N/A	24	
	Physicochemical properties	Presence of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids coded using 11 bits, one bit per property (per residue)	3	3×11=33	40
		Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids (11 properties) within the window	11	7	
	Relative position	Linear distance (in sequence positions) to nearest putative helix, to the nearest putative coil, and to the nearest putative beta sheet	N/A	3	55
		Linear distance (in sequence positions) to nearest conserved residue (conserved = ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	N/A	2	
		Linear distance (in sequence positions) to nearest putative solvent exposed residue (solvent exposed = RSA > 5% or RSA > 20%)	N/A	2	
		Linear distance (in sequence positions) to nearest putative protein-binding disordered residue	N/A	1	
		Linear distance (in sequence positions) to nearest sequence terminus	N/A	1	
		Linear distance (in sequence positions) to nearest residues with high (above average) propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding	N/A	21	
		Linear distance (in sequence positions) to nearest residues that is hydrophobic, positively charged, negatively charged, polar, aliphatic, sulphur containing, aromatic, acidic, small, tiny, and hydroxylic	N/A	11	
		Linear distance (in sequence positions) to nearest secondary structure motifs (X[H|E|C](head), [H|E|C]X(tail), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH)	N/A	14	
	Features that combine multiple structural, physicochemical and evolutionary properties	Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) in the set of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding (3×7=21 values) within the window	11	2×21=42	688
		Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) in the set of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding (3×7=21 values) within the window	11	2×21=42	
		Fraction of putative protein-binding disordered residues in the set of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within the window	11	1×21=21	

Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) in the set of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within the window	11	11×21=231
Fraction of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding located in putative helix/coil/sheet segments within the window	11	3×21=63
Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within the window	11	2×2=4
Fraction of putative protein-binding disordered residues among putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within the window	11	2×1=2
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within the window	11	2×11=22
Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) located in putative helix/coil/sheet segments within the window	11	2×3=6
Fraction of putative protein-binding disordered residues among the conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) within the window	11	2×1=2
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) among the conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) within the window	11	2×11=22
Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) located in putative helix/coil/sheet segments within the window	11	2×3=6
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) among putative protein-binding disordered residues within the window	11	1×11=11
Fraction of putative protein-binding disordered residues in putative helix/coil/sheet segments within the window	11	1×3=3
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) located in putative helix/coil/sheet segments within the window	11	11×3=33
Fraction of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×21=42
Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) located on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×2=4
Fraction of putative helix, coil, and sheet residues on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×3=6
Fraction of putative protein-binding disordered residues on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×1=2
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×11=22
Fraction of residues in different secondary structure motifs (X[H|E|C](head), [H|E|C]X(tail), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH) on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×14=28
Fraction of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within current segment of putative secondary structure	N/A	3×7=21
Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within segment of putative secondary structure that includes the predicted residue	N/A	2
Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) within segment of putative secondary structure that includes the predicted residue	N/A	2
Fraction of putative protein-binding disordered residues within segment of putative secondary structure that includes the predicted residue	N/A	1
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) within segment of putative secondary structure that includes the predicted residue	N/A	11
Fraction of residues with high propensity for {protein; DNA; RNA; small ligands; RNA+DNA+small ligands; max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands); max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein}-binding within motif of putative secondary structure that includes the predicted residue	N/A	3×7=21
Fraction of putative residues surface (two thresholds: RSA > 5% or RSA > 20%) within motif of putative secondary structure that includes the predicted residue	N/A	2
Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) residues within motif of putative secondary structure that includes the predicted residue	N/A	2
Fraction of putative protein-binding disordered residues within motif of putative secondary structure that includes the predicted residue	N/A	1
Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) within motif of putative secondary structure that includes the predicted residue	N/A	11
TOTAL number of features		1090

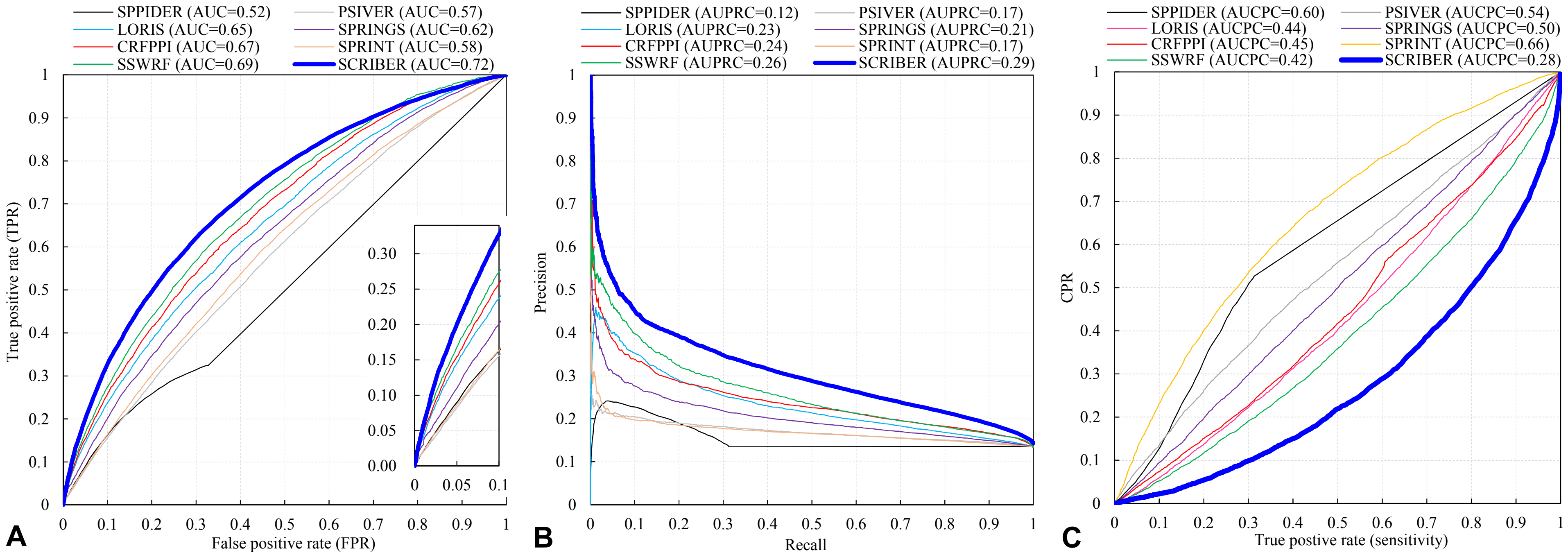
Supplementary Table S3. Description of features that are utilized in the second layer of SCRIBER. Features are computed based on values aggregated over the entire window for window size = 11 (typically average and standard deviations for the 11 values within the window).

Feature type	Description	Window Size	Number of features	Total number of features per feature type
Putative propensity for protein-binding (PB)	Putative PB propensities generated in the first layer (per residue)	11	11	25
	Average of putative PB propensities generated in the first layer	11	1	
	Standard deviation of putative PB propensities generated in the first layer	11	1	
	Fraction of residues with high putative PB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative PB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative PB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative PB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative PB propensities generated in the first layer	11	1	
Putative propensity for DNA-binding (DB)	Putative DB propensities generated in the first layer (per residue)	11	11	25
	Average of putative DB propensities generated in the first layer	11	1	
	Standard deviation of putative DB propensities generated in the first layer	11	1	
	Fraction of residues with high putative DB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative DB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative DB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative DB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative DB propensities generated in the first layer	11	1	
Putative propensity for RNA-binding (RB)	Putative RB propensities generated in the first layer (per residue)	11	11	25
	Average of putative RB propensities generated in the first layer	11	1	
	Standard deviation of putative RB propensities generated in the first layer	11	1	
	Fraction of residues with high putative RB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative RB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative RB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative RB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative RB propensities generated in the first layer	11	1	
Putative propensity for small ligand-binding (SLB)	Putative SLB propensities generated in the first layer (per residue)	11	11	25
	Average of putative SLB propensities generated in the first layer	11	1	
	Standard deviation of putative SLB propensities generated in the first layer	11	1	
	Fraction of residues with high putative SLB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative SLB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative SLB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative SLB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative SLB propensities generated in the first layer	11	1	
Putative propensity for RNA+DNA+small ligands -binding (RDSB)	Putative RDSB propensities generated in the first layer (per residue)	11	11	25
	Average of putative RDSB propensities generated in the first layer	11	1	
	Standard deviation of putative RDSB propensities generated in the first layer	11	1	
	Fraction of residues with high putative RDSB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative RDSB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two	11	2×2=4	

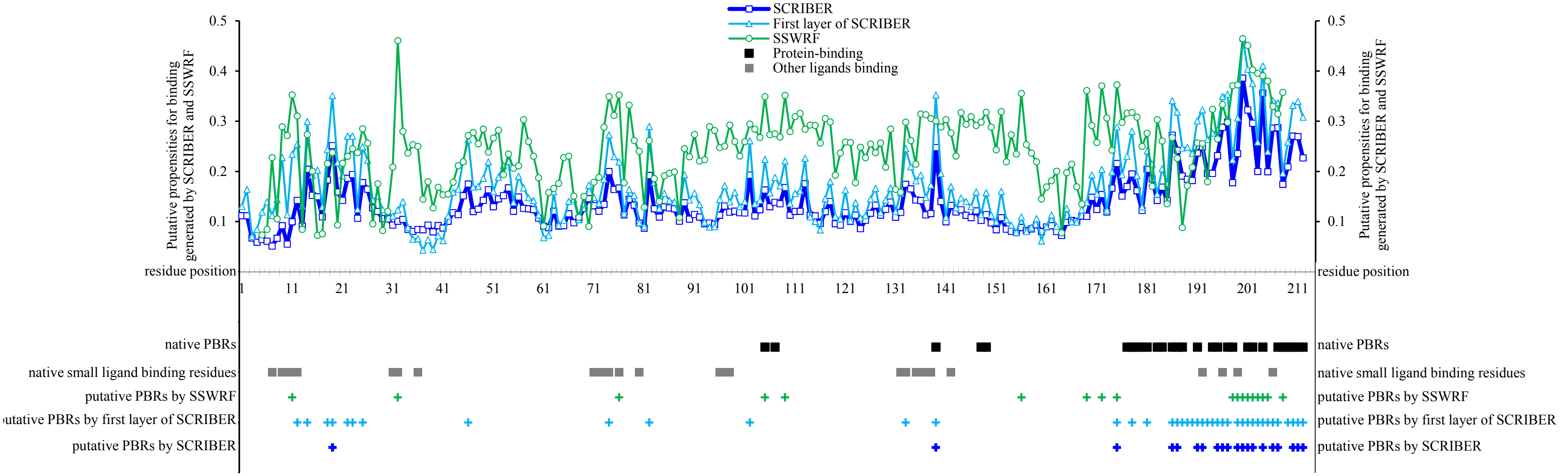
	thresholds: RSA > 5% or RSA > 20%)			
	Fraction of residues with high putative RDSB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative RDSB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative RDSB propensities generated in the first layer	11	1	
Putative max propensity (maxB) = max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands)	Putative maxB propensities generated in the first layer (per residue)	11	11	25
	Average of putative maxB propensities generated in the first layer	11	1	
	Standard deviation of putative maxB propensities generated in the first layer	11	1	
	Fraction of residues with high putative maxB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative maxB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative maxB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative maxB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative maxB propensities generated in the first layer	11	1	
Putative propensity difference (diffB) = max(protein, DNA, RNA, small ligands, RNA+DNA+small ligands) – protein propensity	Putative diffB propensities generated in the first layer (per residue)	11	11	25
	Average of putative diffB propensities generated in the first layer	11	1	
	Standard deviation of putative diffB propensities generated in the first layer	11	1	
	Fraction of residues with high putative diffB propensities generated in the first layer (> average and > weighted average) within the window	11	2	
	Fraction of residues with high putative diffB propensities generated in the first layer (> average and > weighted average) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%)	11	2×2=4	
	Fraction of residues with high putative diffB propensities generated in the first layer (> average and > weighted average) among conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	11	2×2=4	
	Weighted average of putative maxPB propensities generated in the first layer	11	1	
	Weighted standard deviation of putative diffB propensities generated in the first layer	11	1	
	TOTAL number of features			175

Supplementary Table S4. Results of the ablation study that compares predictive performance of SCRIBER with its versions that exclude certain combinations of the three novel design ideas. MODEL1 excludes all three design ideas (no novel features, no combined features and no second layer), i.e., we use the prediction of PBRs generated in the first layer using the remaining features. MODEL2 does not use the combined features and the second layer but uses the novel features, i.e., we use the prediction of PBRs generated in the first layer using all but combined features. MODEL3 does not use the second layer but applies both novel and combined features, i.e., we use the prediction of PBRs generated in the first layer using all features. The binary predictions for all methods are calibrated to allow for direct comparison, such that the number of putative PBRs each method generates equals to the number of the native PBRs. The results are computed over 10 subsets of randomly selected 50% of TEST to assess robustness of the predictions and evaluate statistical significance of differences between SCRIBER and the other methods (section 2.3 gives details). We report the corresponding averages, standard deviations (stdev), and *p*-values; differences with *p*-value < 0.05 are assumed statistically significant and shown in ***bold italics*** font. The best value for each measure of predictive quality are shown with **bold** font.

Predictor		Sensitivity	Specificity	Precision	Accuracy	F1	MCC	AUC	AUPRC	AULC	AULCratio	AUCPC	CPR
MODEL1	average	0.256	0.884	0.255	0.800	0.256	0.140	0.647	0.219	0.017	2.519	0.374	0.133
− No novel features	stdev	±0.009	±0.004	±0.009	±0.006	±0.009	±0.0010	±0.009	±0.008	±0.001	±0.118	±0.015	±0.010
− No combined features − No second layer	*p*-value	***1.69E-11***	***4.14E-06***	***2.18E-11***	***1.10E-06***	***1.85E-11***	***2.64E-11***	***1.94E-10***	***2.99E-11***	***9.35E-07***	***1.92E-10***	***8.47E-11***	***8.17E-04***
MODEL2	average	0.293	0.890	0.292	0.810	0.293	0.183	0.691	0.256	0.018	3.052	0.340	0.137
− No combined features	stdev	±0.011	±0.004	±0.011	±0.007	±0.011	±0.015	±0.013	±0.009	±0.001	±0.198	0.018	±0.013
− No second layer	*p*-value	***1.59E-06***	***6.92E-03***	***1.50E-06***	***3.52E-03***	***1.53E-06***	***3.93E-06***	***1.04E-03***	***5.59E-06***	***2.62E-03***	***7.46E-06***	***4.35E-07***	***5.46E-04***
MODEL3	average	0.324	0.895	0.324	0.819	0.324	0.219	0.712	0.281	**0.020**	3.578	0.313	0.125
− No second layer	stdev	±0.011	±0.004	±0.011	±0.006	±0.011	±0.014	±0.013	±0.011	±0.001	±0.218	±0.015	±0.009
	p-value	*0.12*	*0.56*	*0.14*	*0.47*	*0.13*	*0.15*	*0.62*	*0.27*	*0.51*	*0.21*	***2.64E-04***	***4.0E-02***
SCRIBER	average	**0.334**	**0.896**	**0.332**	**0.821**	**0.333**	**0.230**	**0.715**	**0.287**	**0.020**	**3.725**	**0.282**	**0.116**
	stdev	±0.013	±0.004	±0.013	±0.007	±0.013	±0.016	±0.013	±0.012	±0.001	±0.258	±0.014	±0.009



Supplementary Fig. S1. ROC curves (panel A), precision-recall (PR) curves (panel B), and cross-prediction rate (CPR) curves (panel C) for SCRIBER and the other 7 predictors of PBRs that are evaluated on the TEST dataset. The insert in the right bottom corner of panel A illustrates arguably the most interesting low FPR range of ROC curves that is used to quantify the AULC values.



Supplementary Fig. S2. Case study that illustrates predictions generated by the first and second layer of SCRIBER, and compares them to the native annotations of protein-binding and predictions from SSWRF. This figure shows results for protein-binding F420-dependent NADP reductase (Uniprot ID: O29370). The *x*-axis represents the primary sequence. The dark blue, light blue and green lines at the top show the putative propensities for PBRs generated by SCRIBER, PBRs predictions generated by the first layer of SCRIBER and SSWRF, respectively. The black and grey cubes underneath the *x*-axis represent annotations of native PBRs and other-binding residues. Dark blue, blue and green + markers denote PBRs predicted by SCRIBER, the first layer of SCRIBER and SSWRF, respectively.