

¹Golisano College of Computing and Information Sciences Mkk3671@rit.edu



Datasets

- Protein sequences from EMBL-EBI Reference Proteome
- Interactions from BioGRID database

Dataset	No. of proteins	No. of positive pairs	No. of negative pairs
Yeast	3651	50344	50376
Human	7028	73624	73628

Table 1: Statistics of interactions dataset from BioGRID database

Interpretable Structured Learning with Sparse Gated Sequence Encoder for Protein-Protein Interaction Prediction

Kishan KC¹ Feng Cui²

	Model o	lesc	ripti
 We propose interpretal PPIs using variable le Provides interpreta is computationally Makes accurate PP 	ble deep framework, to ength sequences that able sparsity masks. efficient and scalable. PI predictions.	o mod	el
Step 1: Sequence Enco	der with Bi-GRU		
 Handles variable-length Given a sequence s of lembedding layer project Bidirectional GRU to lembedding of amine h_l = BiGRU 	h sequences. length L with amino acts a_l to vector represent $x_l = \mathbf{W}_e a_l$ learn sequential & control of acids in the sequence $J(x_l) = [\overline{\mathrm{GRU}(x_l)}, \overline{\mathrm{GRU}(x_l)}]$	cids [a ntatio textua es. $\overline{(x_l)}$]	x_1, a_2, \dots n x_l :
Step 2: Sparsity Gating Not all amino acids are Learn sparse mask to f Convert h_l to score p_l :	g e informative for intera- focus only on subsets c $p_1 = W_2$ (tanh(action of imp	s. ortant ε + b ₁))
softmax(p)	sparsemax(p)		
Full support	Sparse but distribute	ed	
$\frac{\exp(\mathbf{p}_i)}{\sum_j \exp(p_j)}$	$\operatorname{argmin}_{\{\mathbf{g} \in \Delta^K\}} \mathbf{g} - \mathbf{g} $	$\mathbf{p} _2^2$	argmin _{
 Step 3: Gaussian representation Proteins interacts with patterns. Such diverse information Protein sequence s is endormalised 	esentation th various proteins on can be reflected in encoded to <i>d</i> -dimension	having the un nal Ga	g divers ncertain aussian
Step 4: Pairwise ranki ➤ Minimize the statistic	ng loss al distance between i	nterac	ting pr
for non-interacting pro	oteins	IIUUI AU	in Pro
\succ Wasserstein distance b	between Gaussian repr	resenta	ation of
E_{ij} = Wasser	rstein distance $(\mathcal{N}(\mu_i, X))$	$\Sigma_i)$, ${\cal N}$	$\left(\mu_{j},\Sigma_{j}\right)$
\succ Employ square-expone	ential loss to learn from	n knor	wn intei
	$\mathcal{L} = \sum_{i} \sum_{(i,j) \in \mathbf{V}^+ (i,j)}$	$\sum_{k \in \mathbf{V}^{-}} ($	$E_{ij}^2 + ex$



R.L.] and the NIH [GM116102 to F.C.]



Results

 \succ Our proposed model shows superior performance than state-of-the-

	Data	Ye AUROC	east AP	Hur AUROC	nan AP
	Profiles	0.891±0.004	0.857 ± 0.007	0.870±0.004	0.835±0.005
	_ sequences Profiles	-0.909 ± 0.003 -0.882+0.003	$-\frac{0.912\pm0.004}{0.888+0.002}$	$-0.8/8\pm0.002$ -0.884 ± 0.003	$ 0.882\pm0.003$ $ 0.893\pm0.004$
Ranking	Sequences	0.902 ± 0.003 0.901 ± 0.002	0.904 ± 0.002	0.881 ± 0.002	0.889 ± 0.001
idom Forest	Profiles	0.908 ± 0.002	0.913 ± 0.003	0.891 ±0.005*	0.896±0.005*
	Sequences		$-\frac{0.925\pm0.001^{*}}{0.995\pm0.005}$	-0.887 ± 0.002	$- 0.894 \pm 0.001$
Ranking	Sequences	0.882 ± 0.006 0.898 ± 0.001	0.885 ± 0.006 0.900 ± 0.002	0.873 ± 0.09 0.874 + 0.002	0.881 ± 0.01 0.883 + 0.001
dom Earast	Profiles	0.906 ± 0.004	0.912 ± 0.002	0.872 ± 0.015	0.877 ± 0.015
uom rorest	Sequences	0.919±0.003	0.921 ± 0.002	0.881 ± 0.002	0.886±0.001
Table 2: (Comparison	with the stat	te-of-the-art m	nodels	
ngmo	chanica	improt	the per	rformona	
	chamsn	n mprov	e the per		
t10n (
Model c	onfiguration	AU	ROC	AP	
No	gating	0.880	± 0.001 0.8	75±0.003	
	Softn	nax 0.881	$\pm 0.001 - 0.8$	$7\bar{7}\pm 0.001$	
oint + RF	Fused	max 0.909	± 0.001 0.9	12 ± 0.002	
-	Sparse	emax 0.913	± 0.001 0.9	16 ± 0.002	
	Softn	nax = 0.882	$\pm 0.001 - 0.8$	79 ± 0.002	
	S 01 (1)	5.002		21 ± 0.001	
ussian + R	F Fused	max 0.919	± 0.003 0.9	$\Delta I \perp U U U U$	
$\frac{1}{1}$	CF Fused Sparse	max 0.919 2 max 0.924	$\pm 0.003 0.92 \\ \pm 0.002 0.92 \\ \hline$	25±0.001	
Table 3: S Sity m	EF Fused Sparse Study of mo	max 0.919 max 0.924 odel componen otch biole	± 0.003 0.92 ± 0.002 0.92 Ints on Yeast d ogical kn	25 ± 0.001 ataset owledge?	
Table 3: S rsity m	F Fused Sparse	max 0.919 max 0.924 odel componen atch biole	± 0.003 0.92 ± 0.002 0.92 Ints on Yeast d ogical kn	25±0.001 ataset owledge?	Model's predic
Table 3: S rsity m	F Fused Sparse Study of mo nask ma	max 0.919 max 0.924 odel componen atch biole	± 0.003 0.92 ± 0.002 0.92 Ints on Yeast d ogical kn	25±0.001 ataset owledge?	Model's predic PF01423 PF14807
Table 3: S Table 3: S rsity m	EF Fused Sparse Study of mo nask ma lignment	max 0.919 max 0.924 odel componer atch biole	± 0.003 0.92 ± 0.002 0.92 Ints on Yeast d ogical kn	25±0.001 25±0.001 .ataset owledge?	Model's predic PF01423 PF14807
ussian + R Table 3: ۶ rsity m م) with	EF Fused Sparse Study of mo nask ma lignment motifs (%)	max 0.919 max 0.924 odel componer atch biole	± 0.003 0.92 ± 0.002 0.92 hts on Yeast d ogical kn HIKVDGECLIASENGEDKNINLEFITNVERNESS F	25±0.001 25±0.001 .ataset owledge?	Model's predic PF01423 PF14807
Table 3: S Table 3: S rsity m Al (5) with	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05	max 0.919 max 0.924 odel componen atch biolo	± 0.003 0.92 ± 0.002 0.92 ats on Yeast d ogical kn HIKVDGECLIASLNGFDKNTNLFITNVFNRIS F	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF01423 PF14807
ussian + R Table 3: ٤ rsity m آه) with	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05 65.63	max 0.919 max 0.924 odel componer atch biole	± 0.003 0.92 ± 0.002 0.92 hts on Yeast d ogical kn F	25±0.001 25±0.001 .ataset owledge?	Model's predic PF01423 PF14807
Table 3: 5 Table 3: 5 rsity m (A) with	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05 65.63	max 0.919 max 0.924 odel componer atch biole	$\pm 0.003 0.92$ $\pm 0.002 0.92$ $nts on Yeast d$ $ogical kn$	25±0.001 25±0.001 .ataset owledge?	Model's predic PF01423 PF14807
Table 3: 5 Table 3: 5 rsity m. (A) (b) with	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componen atch biole	±0.003 0.92 ±0.002 0.92 hts on Yeast d ogical kn F	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYE SKTK Model's predic PF01423 PF1423 PF1423 PF1423 PF14209 WINRERFISKLFLRGDSVIVVLKTPVE
Table 3: § Table 3: § rsity m Al (5) with	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componen atch biolo	±0.003 0.92 ±0.002 0.92 hts on Yeast d ogical kn F	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 WINRERFISKLFLRGDSVIVVLKTPVE
Table 3: 5 Table 3: 5 rsity m Al (5) with	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componer atch biole	$\pm 0.003 0.92$ $\pm 0.002 0.92$ $ats on Yeast defined the second seco$	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 WINRERF I SKLFLRGDSV I VVLKTPVE
Table 3: 5 Table 3: 5 rsity m (A) (A) (A) (A) (A) (A) (A) (A) (A) (A)	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componer atch biole	$\pm 0.003 0.92$ $\pm 0.002 0.92$ nts on Yeast d $Ogical kn$ F $ELEELEEFEFKHGPMSLINDAMVTRTPVIISLRFig$	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV I WEKVYESKTK Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE
Table 3: 5 Table 3: 5 rsity m Al (a) with	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componen otch biolo	±0.003 0.92 ±0.002 0.92 nts on Yeast d ogical kn Gogical kn Fig	25±0.001 25±0.001 ataset owledge? ig 3: LSM8	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 WINRERF I SKLFLRGDSV I VVLKTPVE
Table 3: 5 Table 3: 5 rsity m Al (a) with	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componer otch biolo	±0.003 0.92 ±0.002 0.92 nts on Yeast d ogical kn f cogical kn F f eLEELEEFEFKHGPMSLINDAMVTRTPVIISLR Fig	25±0.001 25±0.001 ataset owledge? ig 3: LSM8	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK PF01423 PF17209 WINRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF01423 PF17209 WINRERF I SKLFLRGDSV I VVLKTPVE
Table 3: 5 Table 3: 5 rsity \mathbf{m} (a) with \mathbf{m} with $\mathbf{time c}$ nces to	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componen otch biolo msstikovinkev msstikovinkev mssoi i depektetsea mssoi i depektetsea mssoi i depektetsea	±0.003 0.92 ±0.002 0.92 hts on Yeast d ogical kn fleeteefeerkhopmstindamvtripvilstr Fig	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNKIENEHVIWEKVYESKTK Model's predic PF01423 PF17209 WINRERFISKLFLRGDSVIVVLKTPVE Model's predic PF01423 PF17209 WINRERFISKLFLRGDSVIVVLKTPVE
Table 3: $(x + R)$ Table 3: $(x + R)$ rsity m Al (x + R) (x + R) rable 3: $(x + R)$ rsity m Al (x + R) (x	EF Fused Sparse Study of mo nask ma lignment motifs (%) 59.05 65.63 h motifs	max 0.919 max 0.924 odel componen otch biolo misatekovenkev misatekoven misatekovenkev misatekovenkev misatekovenkev misatekov	±0.003 0.92 ±0.002 0.92 nts on Yeast d oggical kn f cogical kn F ELEELEEFEFKHGPMSLINDAMVTRTPVIISIR Fig LACRSCPYEFFIEGIEIYDRKKLPRKEVDDVL Fig t	25±0.001 25±0.001 ataset owledge? owledge? ig 3: LSM8 NNHKIIARVKAFDRHCMMVLENVKELMTEKKGKI 4: SMD2	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE
Table 3: 5 Table 3: 5 rsity m Al (a) with e mask with time c nces to sed on	EF Fused Sparse Study of mo hask ma hask ma lignment motifs (%) 59.05 65.63 h motifs b motifs	max 0.919 max 0.924 odel componen atch biole	±0.003 0.92 ±0.002 0.92 nts on Yeast d ogical kn fleeteeteeferkhopmstindamvtriteviiiste Fig tation ions	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNKTENEHVIWEKVYESKTK Model's predic PF01423 PF17209 WINRERFISKLFLRGDSVIVVLKTPVE Model's predic PF13248 PF17209 WINRERFISKLFLRGDSVIVVLKTPVE Model's predic PF13248 PF17209 F01485 PF12760 PF01485 PF12760 PF01495 PF0195 PF01495 PF01495 PF01495 PF01495 PF01495 PF01495 PF0195
Table 3: 5 Table 3: 5 rsity m (A) (a) with (b) with (c) c) (c) c) c) (c) c) c) (c) c) c) (c) c) c) (c) c) c	EF Fused Sparse Study of mo ask ma hask ma ignment motifs (%) 59.05 65.63 h motifs h motifs	max 0.919 max 0.924 odel componer otch biolo msatikovinkev mssqi i drekkelska mssqi i drekkelska son cepresent interact	±0.003 0.92 ±0.002 0.92 hts on Yeast d ogical kn ogical kn F f f f f f f f f f f f f f f f f f f	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF12760 PF01096 RSADEPMTTFYKCVNCGHRWKEN
Table 3: 5 Table 3: 5 rsity m (a) with (b) with (c) mask w	EF Fused Sparse Study of mo hask ma hask ma lignment motifs (%) 59.05 65.63 h motifs b motifs compari b their is known PIPR)	max 0.919 max 0.924 odel componen atch biole	±0.003 0.92 ±0.002 0.92 hts on Yeast d ogical kn ogical kn f trodeccelasengedkataleftenverse F	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV I WEKVYESKTK Model's predic PF01423 PF17209 W I NRERF I SKLFLRGDSV I VVLKTPVE Model's predic PF13248 PF13248 PF13248 PF13268 PF1758 PF1
Table 3: 5 Table 3: 5 rsity m (A) (a) with (b) with (c) mask mith (c) mask mith) (c) mas	EF Fused Sparse Study of mo ask ma lignment motifs (%) 59.05 65.63 h motifs b their is compari o their is known PIPR) scalabl	max 0.919 max 0.924 odel component atch biolo missor looperteesent interactions encodes le to lar	±0.003 0.92 ±0.002 0.92 nts on Yeast d ogical kn ogical kn f teleeleefefkhophslindamytripviliste Fig tation ions pairs ge	25±0.001 25±0.001 ataset owledge?	Wodel's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWERVYESKTK Model's predic PF01423 PF17209 W I NRERF I SKLF LRGDSV I VVLKTPVE Model's predic PF13248 PF13248 PF13248 PF13248 PF13248 PF13248 PF13248 PF13248 PF13248 PF13248 PF13260 PF13260 PF13260 PF13260 PF13261 PF1
Table 3: 5 Table 3: 5 rsity m (a) with (b) with (c) with (c) mask with (EF Fused Sparse Study of mo hask ma hask ma lignment motifs (%) 59.05 65.63 h motifs h motifs compari h heir i known PIPR) scalabl	max 0.919 max 0.924 odel componer otch biole msstekovener msstekovener mssoi i drevener mssoi drevener mssoi i drevener mssoi i drevener mssoi i drevener msso	±0.003 0.92 ±0.002 0.92 nts on Yeast d ogical kn ogical kn f treeeleefeerkhoppist indamytripyi istr Fig tation ions pairs ge	25±0.001 25±0.001 ataset owledge?	Model's predic PF01423 PF14807 DEKKVPMLKDTKNK I ENEHV IWEKVYESKTK Model's predic PF01423 PF17209 WINRERF I SKLFLRGDSV I WULKTPVE Model's predic PF01423 PF17209 WINRERF I SKLFLRGDSV I WULKTPVE Model's predic PF01296 PF01096 SADEPMTTFYKCVNCGHRWKEN PF01096 SADEPMTTFYKCVNCGHRWKEN

References

Hashemifar, Somaye, et al. "Predicting protein-protein interactions through sequence-based deep learning." *Bioinformatics* 34.17 (2018): i802-i810. 2. Chen, Muhao, et al. "Multifaceted protein-protein interaction prediction based on Siamese residual rcnn." *Bioinformatics*35.14 (2019): i305-i314.

Acknowledgements

This work was supported by the NSF [NSF-1062422 to A.H.], [NSF-1850492 to