TEXAS A&M UNIVERSITY
Engineering

3DSIG COSI

Cross-Modality and Self-Supervised Protein Embedding for Compound-Protein Affinity and Contact Prediction

Yuning You and Yang Shen
Texas A&M University

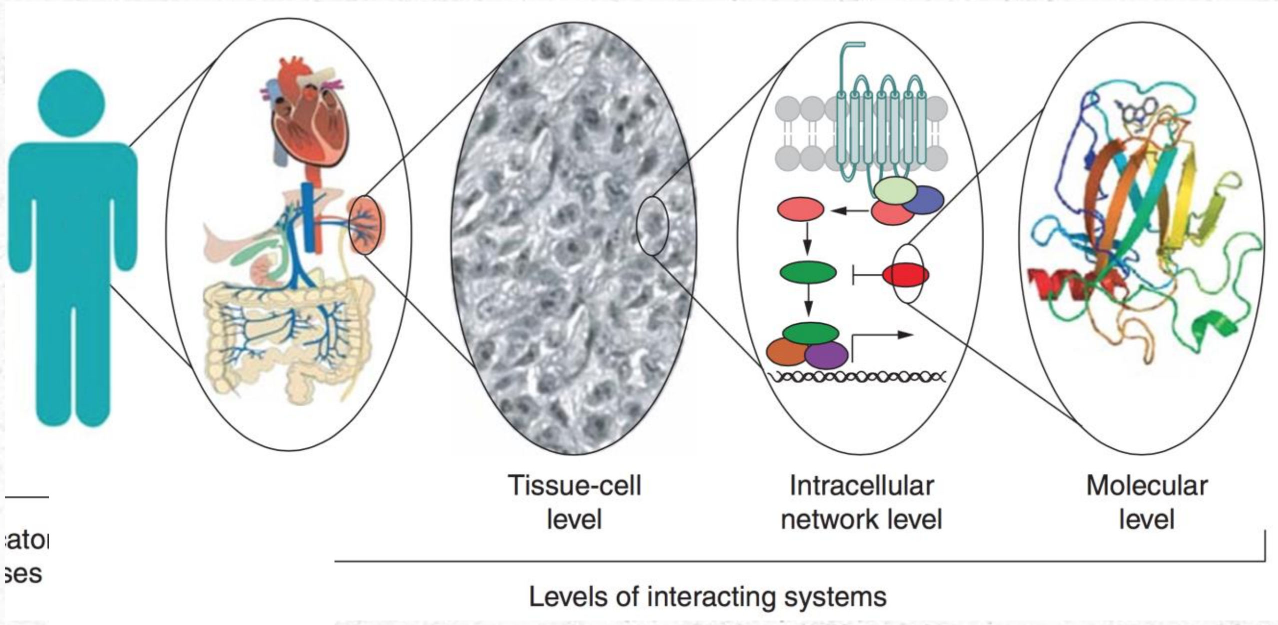
ISMB/ECCB 2021 – 3DSIG

July 25, 2021

Motivation: Drug Discovery for Biological Systems

- A paradigm shift

  ~~One disease. One target. One drug~~ => **Systems Pharmacology**.

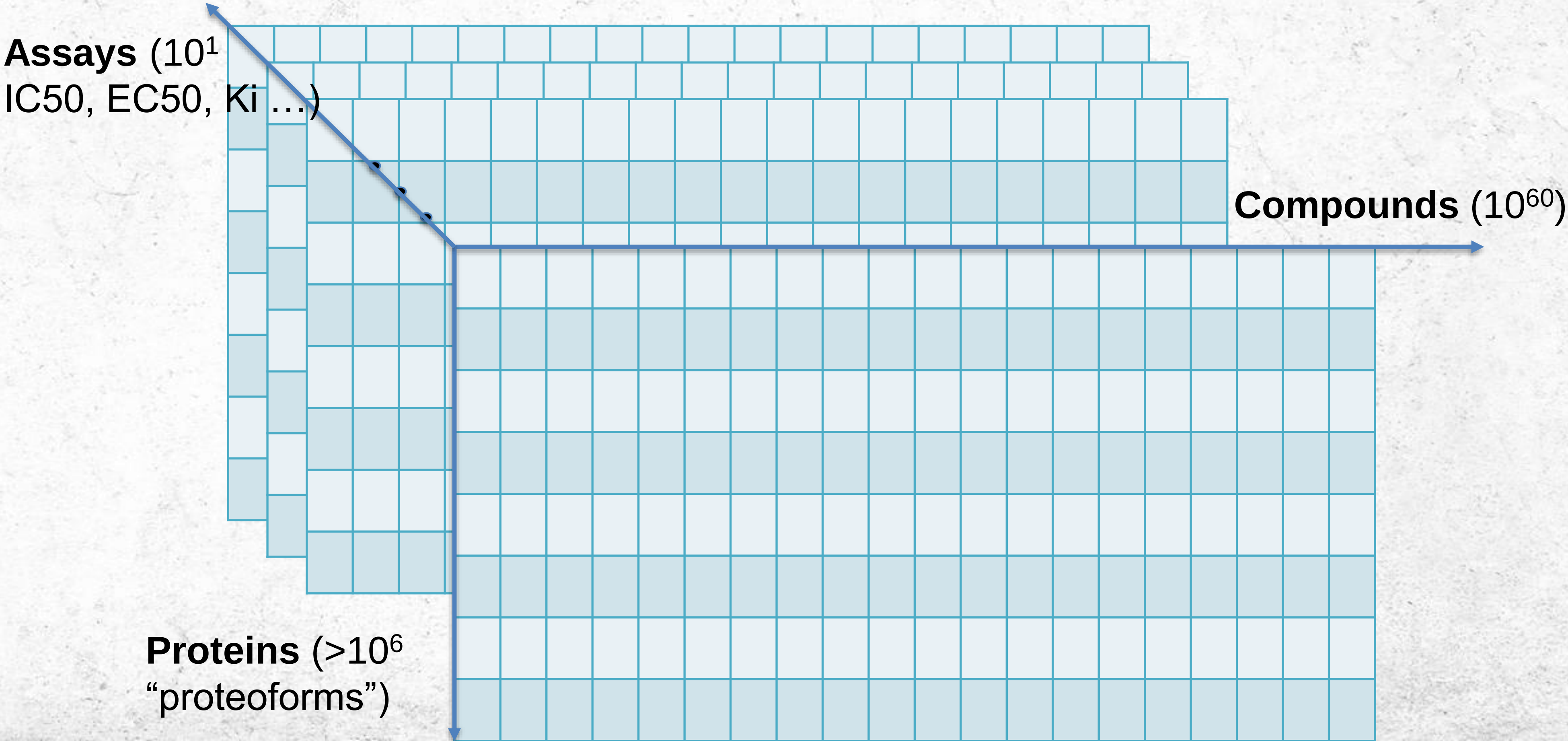


  - Desired multiple targets (with proper activity profiles)
  - Undesired multiple targets to avoid toxicity and side-effects.

Wist et. al. *Genome Medicine* (2009)

Exploring Chemical Space for Desired Interactions on Proteomes

Over 80% of >900 FDA-approved human drugs are small-molecule compounds targeting proteins



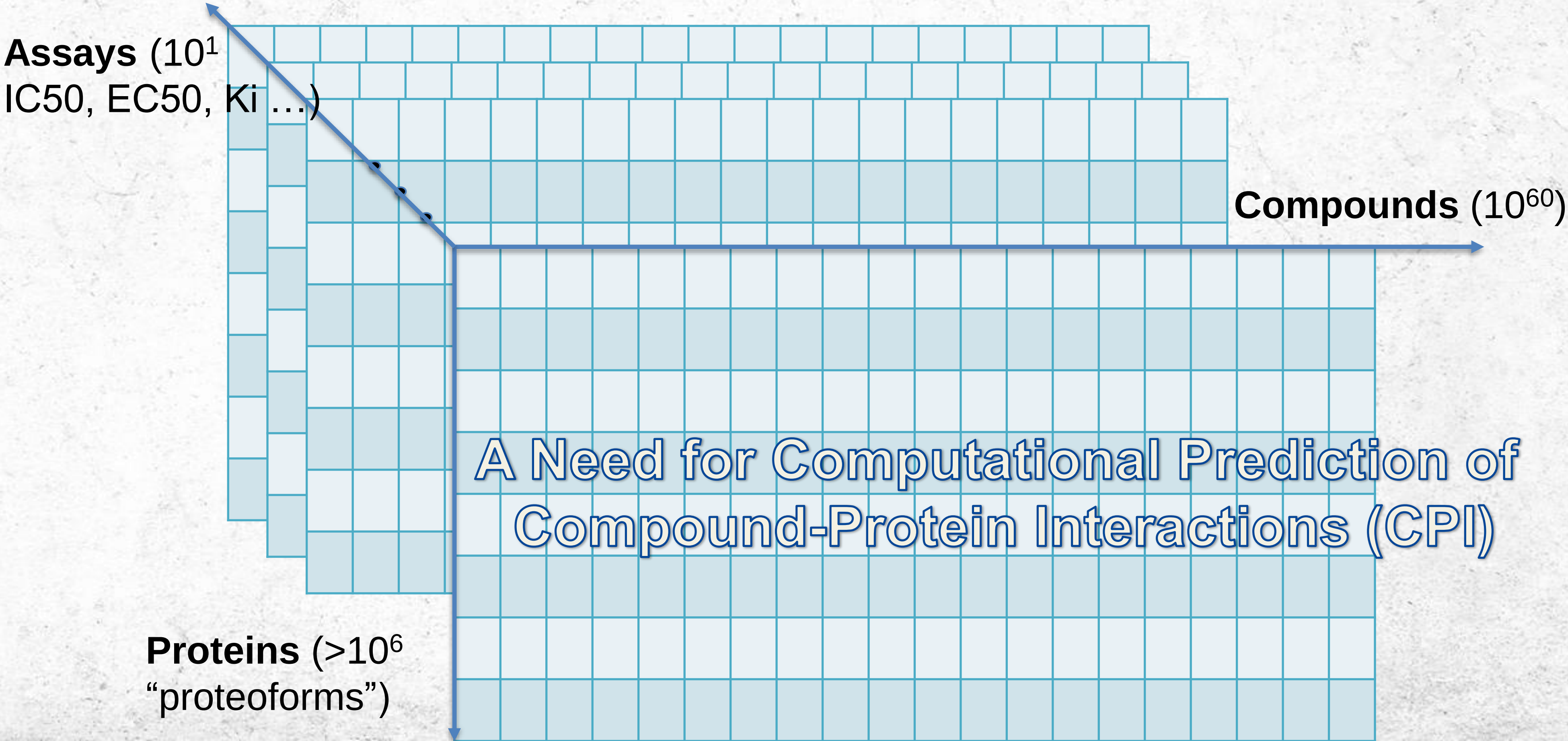
Assays (10^1 IC50, EC50, Ki …)

Compounds (10^{60})

Proteins ($>10^6$ “proteoforms”)

Exploring Chemical Space for Desired Interactions on Proteomes

Over 80% of >900 FDA-approved human drugs are small-molecule compounds targeting proteins

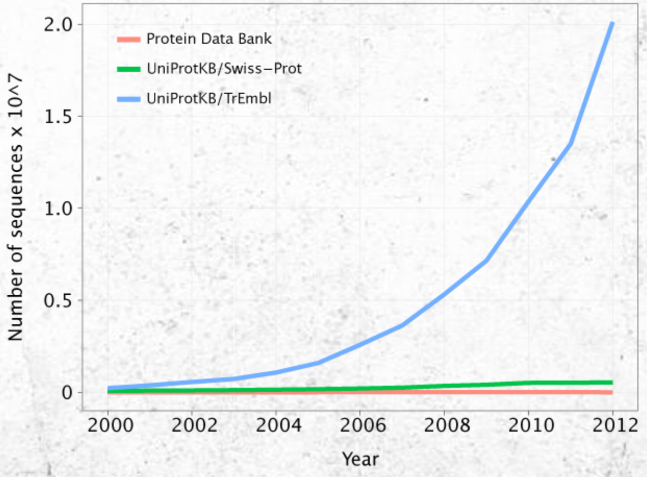


A Need for Computational Prediction of Compound-Protein Interactions (CPI)

Current CPI Prediction Methods

Protein structure-based docking

- Can predict the activity level of CPI (affinity)
- Very interpretable
- Non-convex optimization is challenging and slow
- Many proteins' structures are not solved



Sequenced

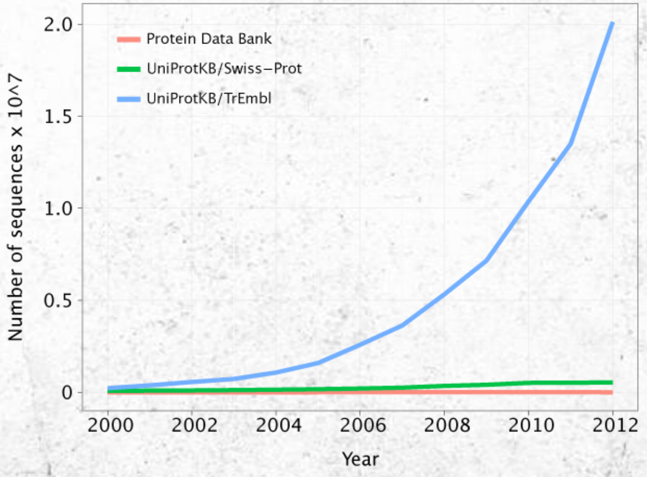
Functionally annotated

Structurally determined

Current CPI Prediction Methods

Protein sequence-based CPI identification (As of 2017)

- Can only classify CPIs (mostly binary)
- Not interpretable
- Machine learning is relatively fast
- Labeled sequence data are abundant



Sequenced

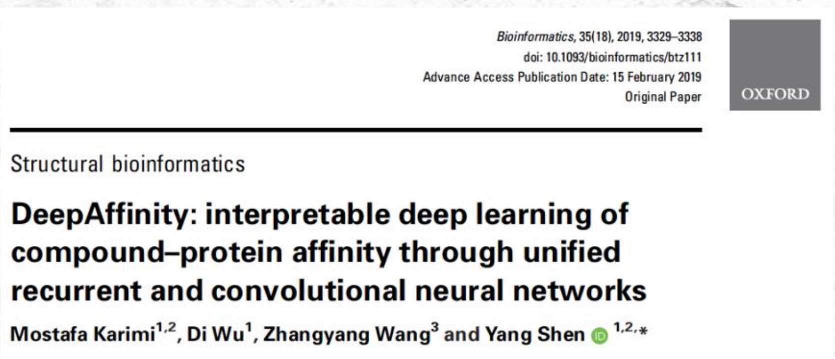
Functionally annotated

Structurally determined

Current CPI Prediction Methods

Protein sequence-based CPI identification (As of 2021)

- Can predict affinity levels
- Somewhat interpretable (“attentions” over intermolecular contacts)
- Machine learning is relatively fast
- Labeled (and unlabeled) sequence data are abundant



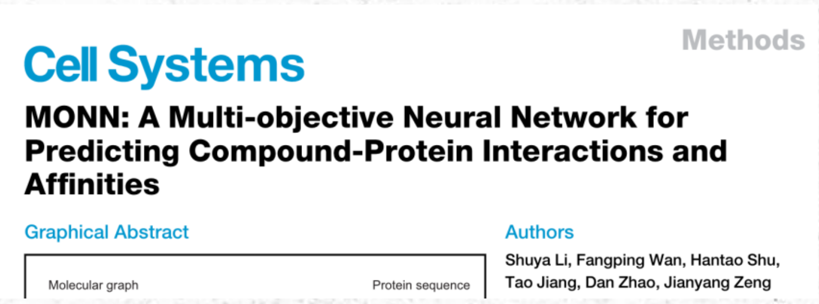
Bioinformatics, 35(18), 2019, 3329–3338
doi: 10.1093/bioinformatics/btz111
Advance Access Publication Date: 15 February 2019
Original Paper

OXFORD

Structural bioinformatics

DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks

Mostafa Karimi[1,2], Di Wu[1], Zhangyang Wang[3] and Yang Shen[1,2,*]



Methods

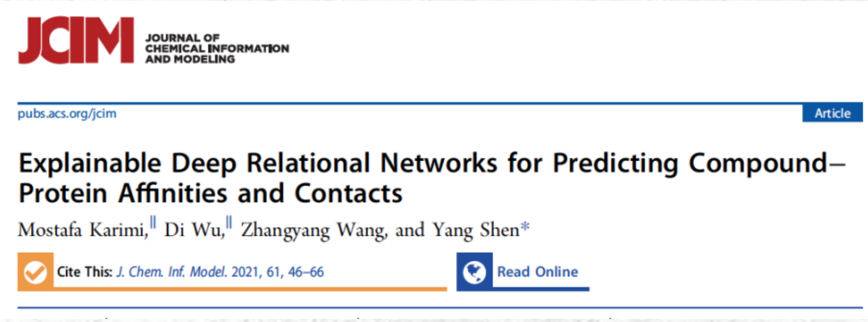
Cell Systems

MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities

Graphical Abstract

Authors

Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, Jianyang Zeng



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

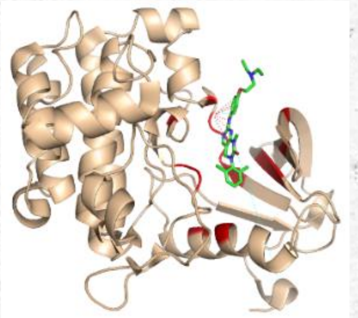
Explainable Deep Relational Networks for Predicting Compound–Protein Affinities and Contacts

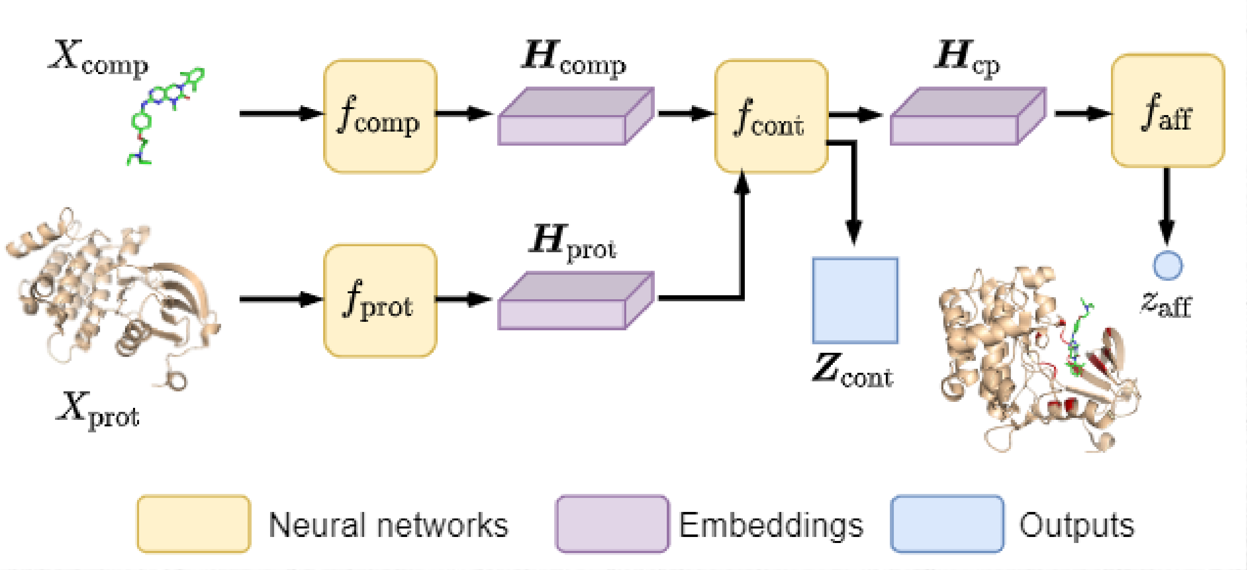
Mostafa Karimi,[‖] Di Wu,[‖] Zhangyang Wang, and Yang Shen*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 46−66 Read Online

Current Formulation

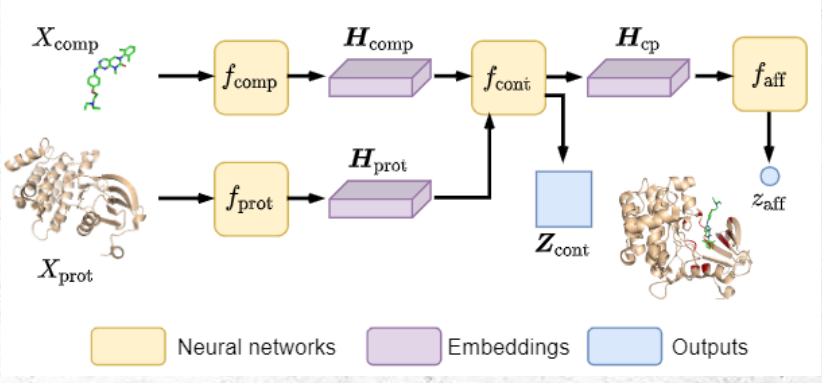
- Compound-protein affinity and contact prediction (CPAC):
  - Affinity: quantitative level of interaction
  - Contact: intermolecular atom-residue contact, underlying ***interpretation*** for affinity





Remaining Gaps

- Structure-relevant prediction relies on structure-unaware 1D sequences as inputs*
  - Not suffice to model 3D structural relationships
  - Empirically less generalizable



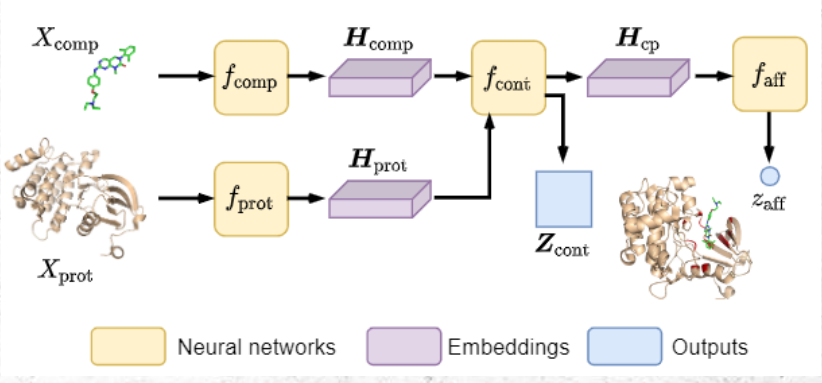
- More severe under sparse ground-truth labeling
  - Pairwise (compound-protein) labels are expensive (especially contact label)
  - Structure data are less available
  - Intersection of them ⇒ Supervision starvation

- Challenges: Inadequate data information & label supervision

* Exceptions: DeepAffinity uses sequence-predicted *structure property sequence* as inputs.
DeepAffinity+/DeepRelations uses sequence-predicted structure contexts as regularization.

Remaining Gaps

- Structure-relevant prediction relies on structure-unaware 1D sequences as inputs*
  - Not suffice to model 3D structural relationships
  - Empirically less generalizable



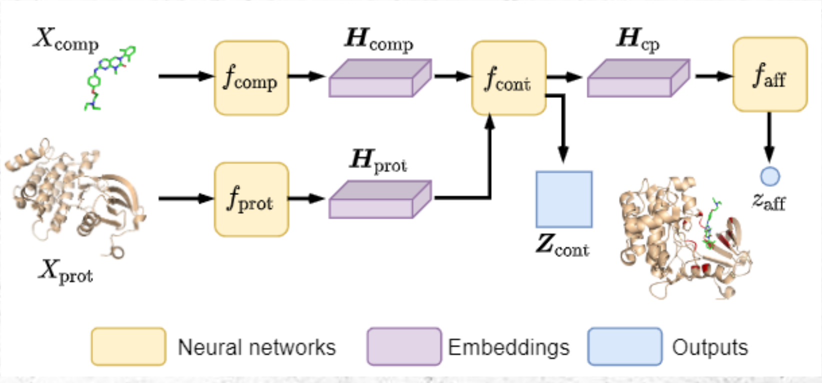
- More severe under sparse ground-truth labeling
  - Pairwise (compound-protein) labels are expensive (especially contact label)
  - Structure data are less available
  - Intersection of them → supervision starvation

- Challenges: Inadequate data information & label supervision

* Exceptions: DeepAffinity uses sequence-predicted *structure property sequence* as inputs.
DeepAffinity+/DeepRelations uses sequence-predicted structure contexts as regularization.

Remaining Gaps

- Structure-relevant prediction relies on structure-unaware 1D sequences as inputs*
  - Not suffice to model 3D structural relationships
  - Empirically less generalizable



- More severe under sparse ground-truth labeling
  - Pairwise (compound-protein) labels are expensive (especially contact label)
  - Structure data are less available
  - Intersection of them → supervision starvation

- Challenges: Inadequate data information & label supervision

* Exceptions: DeepAffinity uses sequence-predicted *structure property sequence* as inputs.
DeepAffinity+/DeepRelations uses sequence-predicted structure contexts as regularization.

Our Contributions

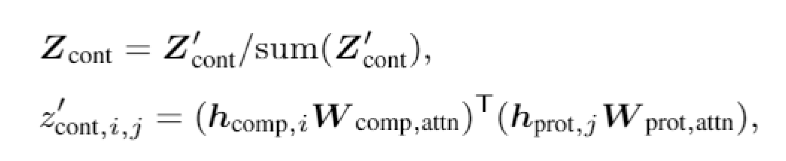
- Cross-modality learning to introduce structure-awareness
  - Different modalities excel at different tasks
  - Concatenation, cross interaction further benefit

- Self-supervised learning to exploit unlabelled data
  - Mask language modeling for 1D model
  - Graph completion for 2D model
  - Different self-supervisions boost different tasks

Ref 6.
arXiv:2012.00651
(MLSB’20)

Method. Cross-Modality Learning

- Base model: DeepAffinity+ Ref 1. DeepRelations, JCIM’20
  - Replace hierarchical attention with joint attention



$$\boldsymbol{Z}_{\text{cont}} = \boldsymbol{Z}'_{\text{cont}} / \text{sum}(\boldsymbol{Z}'_{\text{cont}}),$$

$$z'_{\text{cont},i,j} = (\boldsymbol{h}_{\text{comp},i} \boldsymbol{W}_{\text{comp,attn}})^{\mathsf{T}} (\boldsymbol{h}_{\text{prot},j} \boldsymbol{W}_{\text{prot,attn}}),$$

- Compounds are represented as chemical graphs and encoded by GCN

- Single-modality model for proteins
  - 1D sequence model:
    - Amino-acid sequence (consecutive *k*-mers) as protein input
    - HRNN as sequence encoder
  - 2D graph model:
    - Predicted intra-protein contact map as protein input Ref 2. RaptorX, NAR’16
    - Still structure-free input, with additional structural and evolutional information as induction bias from RaptorX
    - GAT as graph encoder

Method. Cross-Modality Learning

- Cross-modality model:
  - Concatenation
    - Concatenating embeddings of different modalities
    - Preserving information
  - Cross interaction
    - Additional information flow is introduced across modalities

$$\begin{aligned} \boldsymbol{h}_{\text{prot,seq},n} &= \left(\text{sigmoid}(\boldsymbol{h}''_{\text{prot,graph},n}{}^{\mathsf{T}}\boldsymbol{h}'_{\text{prot,seq},n}) + 1\right)\boldsymbol{h}'_{\text{prot,seq},n}, \\ \boldsymbol{h}_{\text{prot,graph},n} &= \left(\text{sigmoid}(\boldsymbol{h}''_{\text{prot,seq},n}{}^{\mathsf{T}}\boldsymbol{h}'_{\text{prot,graph},n}) + 1\right)\boldsymbol{h}'_{\text{prot,seq},n}, \end{aligned} \tag{4}$$

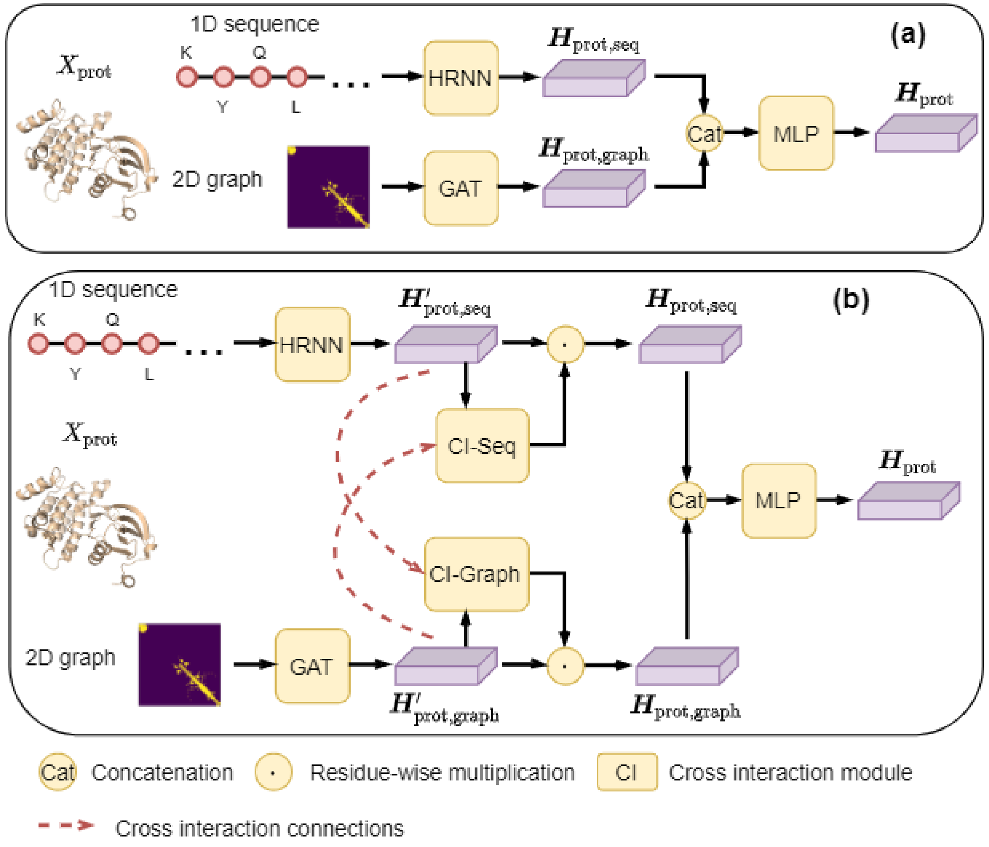


Figure 1: Cross-modality encoders. (a) Naïve concatenation preserves information from different sources. (b) Cross interaction with inter-modality information flows.

Method. Self-Supervised Learning

- Masked language modeling (MLM) for 1D sequences
  - Predicting the masking residue with sequential relation

$$\min_{\{\text{HRNN, MLP}\}} \quad \mathcal{L}_{\text{CE}}\Big(\text{MLP}(\text{HRNN}(\bar{\boldsymbol{F}}_{\text{prot}})), \boldsymbol{Y}_{\text{mask}}\Big),$$

$$\text{s.t.} \quad \bar{\boldsymbol{F}}_{\text{prot}}, \boldsymbol{Y}_{\text{mask}} = \text{mask}(\boldsymbol{F}_{\text{prot}}),$$

- Graph completion (Graph Comp.) for 2D contact maps
  - Predicting the masking residue with topological knowledge

$$\min_{\{\text{GAT, MLP}\}} \quad \mathcal{L}_{\text{CE}}\Big(\text{MLP}(\text{GAT}(\bar{\boldsymbol{F}}_{\text{prot}}, \boldsymbol{A}_{\text{prot}})), \boldsymbol{Y}_{\text{mask}}\Big),$$

$$\text{s.t.} \quad \bar{\boldsymbol{F}}_{\text{prot}}, \boldsymbol{Y}_{\text{mask}} = \text{mask}(\boldsymbol{F}_{\text{prot}}).$$

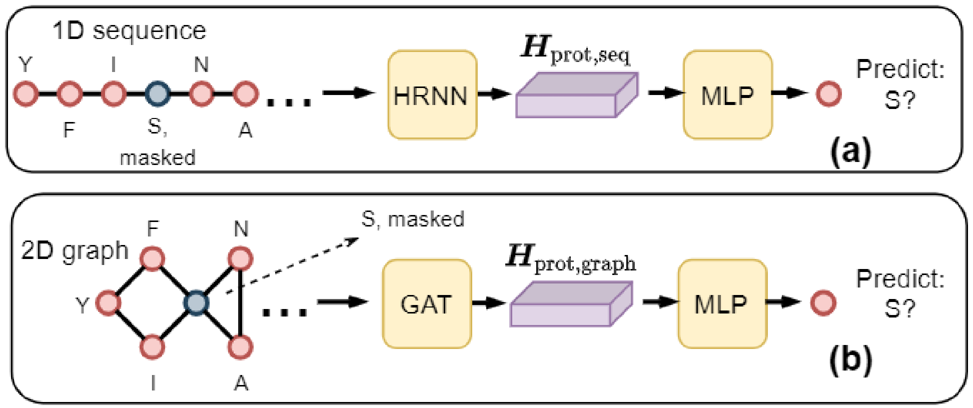


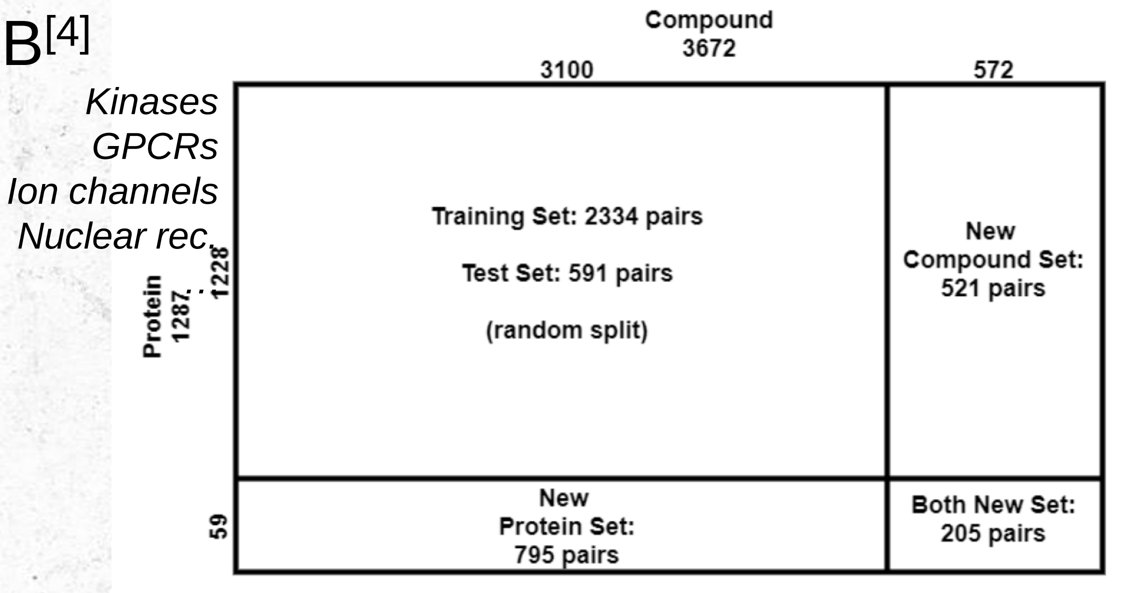
Figure 2: Self-supervised tasks for different modalities. (a) Masked language modeling (MLM). (b) Graph completion (GraphComp).

- Joint pre-training
  - Jointly performing MLM and GraphComp

Experiments. Datasets

- Evaluation dataset
  - CPAC data (~4,500 pairs) from DeepRelations
  - Curated from PDBsum[3] and BindingDB[4]

Ref 1. DeepRelations, JCIM’20



Kinases
GPCRs
Ion channels
Nuclear rec.

Training Set: 2334 pairs

Test Set: 591 pairs

(random split)

New Compound Set: 521 pairs

New Protein Set: 795 pairs

Both New Set: 205 pairs

- Self-supervised pre-training dataset for embedding
  - 12,798,671 protein domain sequences
  - 60,137 sequences with structure
  - Curated from Pfam-A[5]

Experiments. Results

- Single-modality: Different mods. excel at different tasks

❖ 1D seq. in affinity prediction

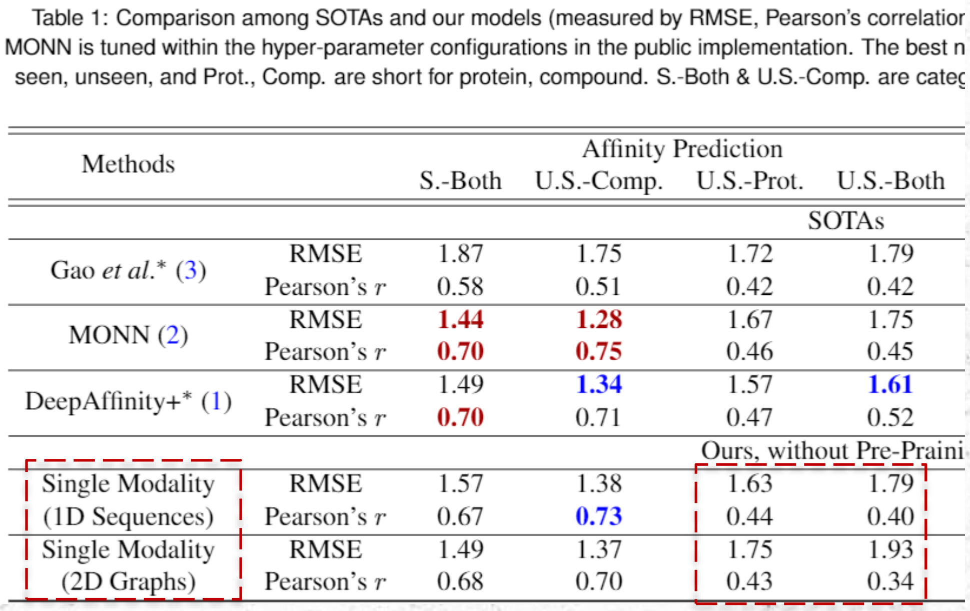


Table 1: Comparison among SOTAs and our models (measured by RMSE, Pearson's correlatior
MONN is tuned within the hyper-parameter configurations in the public implementation. The best n
seen, unseen, and Prot., Comp. are short for protein, compound. S.-Both & U.S.-Comp. are categ

Methods		Affinity Prediction S.-Both	U.S.-Comp.	U.S.-Prot.	U.S.-Both
				SOTAs	
Gao *et al.** (3)	RMSE	1.87	1.75	1.72	1.79
	Pearson's r	0.58	0.51	0.42	0.42
MONN (2)	RMSE	**1.44**	**1.28**	1.67	1.75
	Pearson's r	**0.70**	**0.75**	0.46	0.45
DeepAffinity+* (1)	RMSE	1.49	**1.34**	1.57	**1.61**
	Pearson's r	**0.70**	0.71	0.47	0.52
				Ours, without Pre-Praini	
Single Modality (1D Sequences)	RMSE	1.57	1.38	1.63	1.79
	Pearson's r	0.67	**0.73**	0.44	0.40
Single Modality (2D Graphs)	RMSE	1.49	1.37	1.75	1.93
	Pearson's r	0.68	0.70	0.43	0.34

❖ 2D graph in contact prediction

Experiments. Results

- ➢ Single-modality: Different mods. excel at different tasks
- ❖ 1D seq. in affinity prediction
- ❖ 2D graph in contact prediction

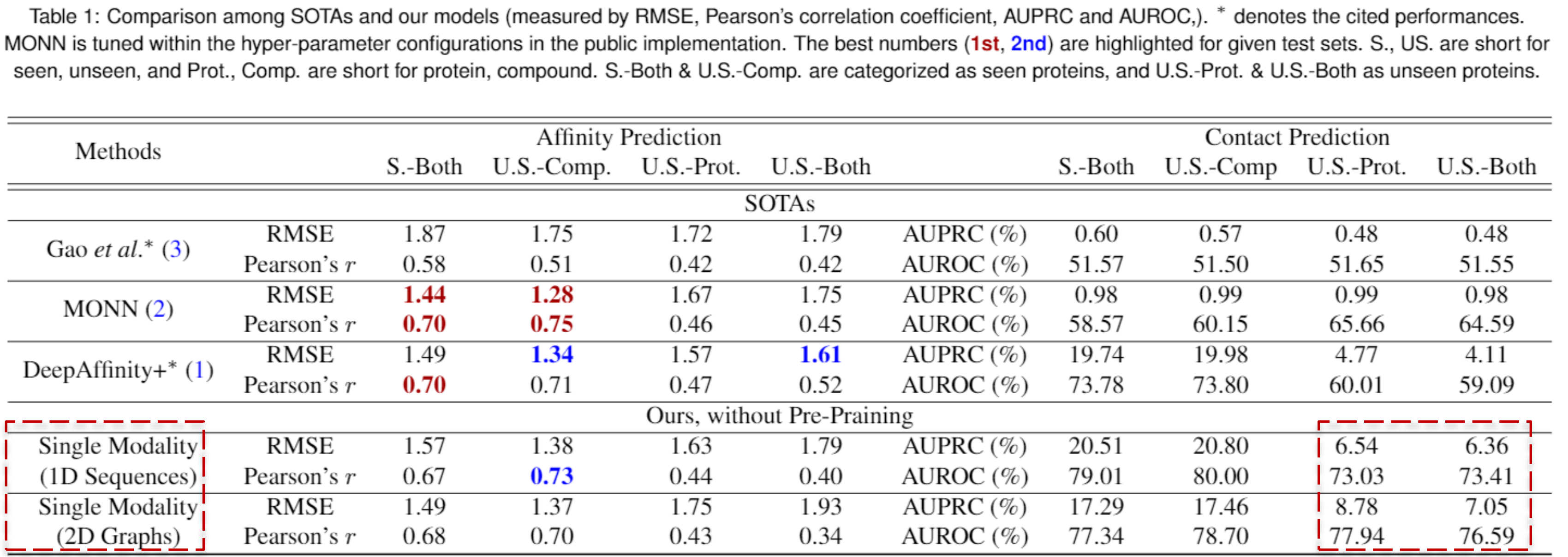


Table 1: Comparison among SOTAs and our models (measured by RMSE, Pearson's correlation coefficient, AUPRC and AUROC,). * denotes the cited performances. MONN is tuned within the hyper-parameter configurations in the public implementation. The best numbers (1st, 2nd) are highlighted for given test sets. S., US. are short for seen, unseen, and Prot., Comp. are short for protein, compound. S.-Both & U.S.-Comp. are categorized as seen proteins, and U.S.-Prot. & U.S.-Both as unseen proteins.

Methods		Affinity Prediction					Contact Prediction			
		S.-Both	U.S.-Comp.	U.S.-Prot.	U.S.-Both		S.-Both	U.S.-Comp	U.S.-Prot.	U.S.-Both
SOTAs										
Gao *et al.** (3)	RMSE	1.87	1.75	1.72	1.79	AUPRC (%)	0.60	0.57	0.48	0.48
	Pearson's r	0.58	0.51	0.42	0.42	AUROC (%)	51.57	51.50	51.65	51.55
MONN (2)	RMSE	**1.44**	**1.28**	1.67	1.75	AUPRC (%)	0.98	0.99	0.99	0.98
	Pearson's r	**0.70**	**0.75**	0.46	0.45	AUROC (%)	58.57	60.15	65.66	64.59
DeepAffinity+* (1)	RMSE	1.49	**1.34**	1.57	**1.61**	AUPRC (%)	19.74	19.98	4.77	4.11
	Pearson's r	**0.70**	0.71	0.47	0.52	AUROC (%)	73.78	73.80	60.01	59.09
Ours, without Pre-Praining										
Single Modality (1D Sequences)	RMSE	1.57	1.38	1.63	1.79	AUPRC (%)	20.51	20.80	6.54	6.36
	Pearson's r	0.67	**0.73**	0.44	0.40	AUROC (%)	79.01	80.00	73.03	73.41
Single Modality (2D Graphs)	RMSE	1.49	1.37	1.75	1.93	AUPRC (%)	17.29	17.46	8.78	7.05
	Pearson's r	0.68	0.70	0.43	0.34	AUROC (%)	77.34	78.70	77.94	76.59

Experiments. Results

- Cross-modality further benefits

- Simple concat. boosts against either mods.

- Further inter-mod. information flow (cross interaction) achieves SOTA

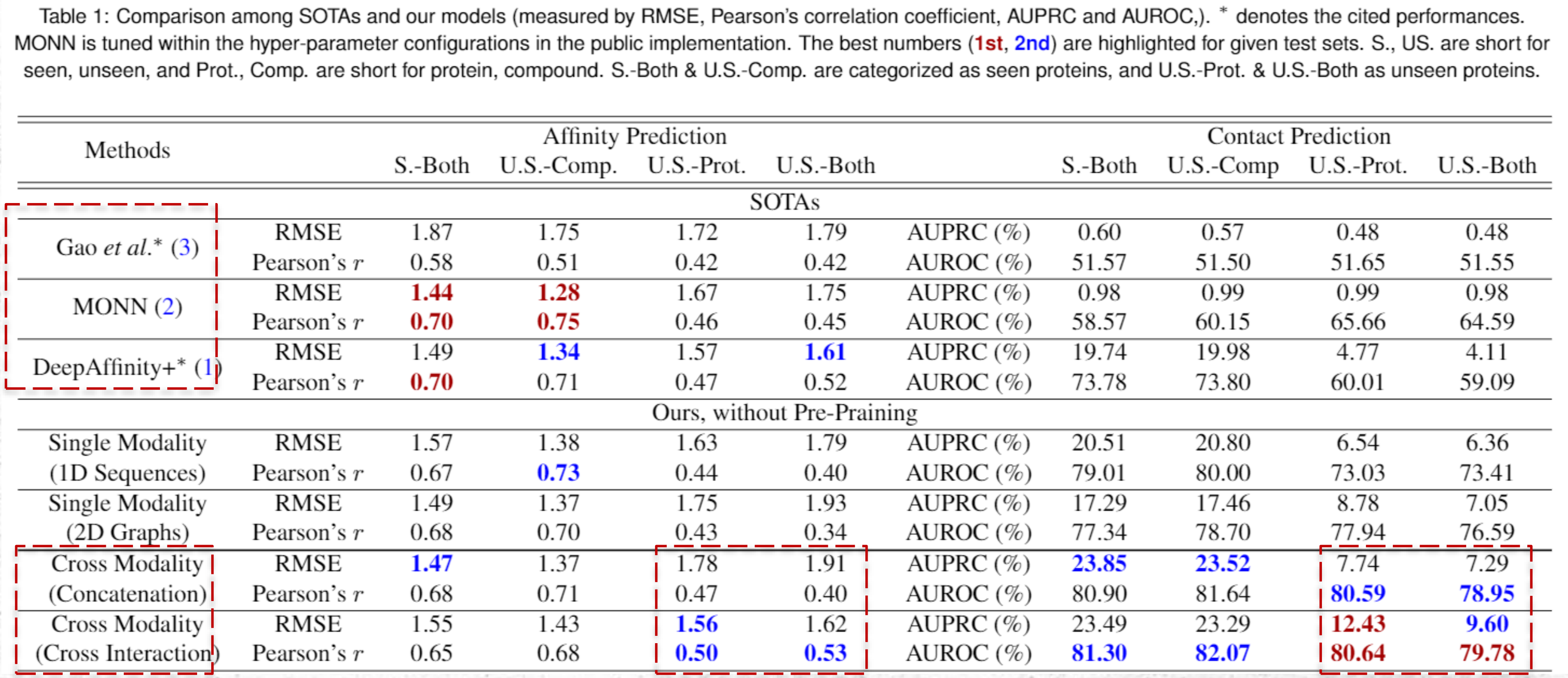


Table 1: Comparison among SOTAs and our models (measured by RMSE, Pearson's correlation coefficient, AUPRC and AUROC,). * denotes the cited performances. MONN is tuned within the hyper-parameter configurations in the public implementation. The best numbers (1st, 2nd) are highlighted for given test sets. S., US. are short for seen, unseen, and Prot., Comp. are short for protein, compound. S.-Both & U.S.-Comp. are categorized as seen proteins, and U.S.-Prot. & U.S.-Both as unseen proteins.

Methods		Affinity Prediction					Contact Prediction			
		S.-Both	U.S.-Comp.	U.S.-Prot.	U.S.-Both		S.-Both	U.S.-Comp	U.S.-Prot.	U.S.-Both
SOTAs										
Gao *et al.** (3)	RMSE	1.87	1.75	1.72	1.79	AUPRC (%)	0.60	0.57	0.48	0.48
	Pearson's r	0.58	0.51	0.42	0.42	AUROC (%)	51.57	51.50	51.65	51.55
MONN (2)	RMSE	**1.44**	**1.28**	1.67	1.75	AUPRC (%)	0.98	0.99	0.99	0.98
	Pearson's r	**0.70**	**0.75**	0.46	0.45	AUROC (%)	58.57	60.15	65.66	64.59
DeepAffinity+* (1)	RMSE	1.49	**1.34**	1.57	**1.61**	AUPRC (%)	19.74	19.98	4.77	4.11
	Pearson's r	**0.70**	0.71	0.47	0.52	AUROC (%)	73.78	73.80	60.01	59.09
Ours, without Pre-Praining										
Single Modality (1D Sequences)	RMSE	1.57	1.38	1.63	1.79	AUPRC (%)	20.51	20.80	6.54	6.36
	Pearson's r	0.67	**0.73**	0.44	0.40	AUROC (%)	79.01	80.00	73.03	73.41
Single Modality (2D Graphs)	RMSE	1.49	1.37	1.75	1.93	AUPRC (%)	17.29	17.46	8.78	7.05
	Pearson's r	0.68	0.70	0.43	0.34	AUROC (%)	77.34	78.70	77.94	76.59
Cross Modality (Concatenation)	RMSE	**1.47**	1.37	1.78	1.91	AUPRC (%)	**23.85**	**23.52**	7.74	7.29
	Pearson's r	0.68	0.71	0.47	0.40	AUROC (%)	80.90	81.64	**80.59**	**78.95**
Cross Modality (Cross Interaction)	RMSE	1.55	1.43	**1.56**	1.62	AUPRC (%)	23.49	23.29	**12.43**	**9.60**
	Pearson's r	0.65	0.68	**0.50**	**0.53**	AUROC (%)	**81.30**	**82.07**	**80.64**	**79.78**

Experiments. Results

- 1D pre-training (MLM) promotes affinity prediction

- Deteriorating contact prediction performance

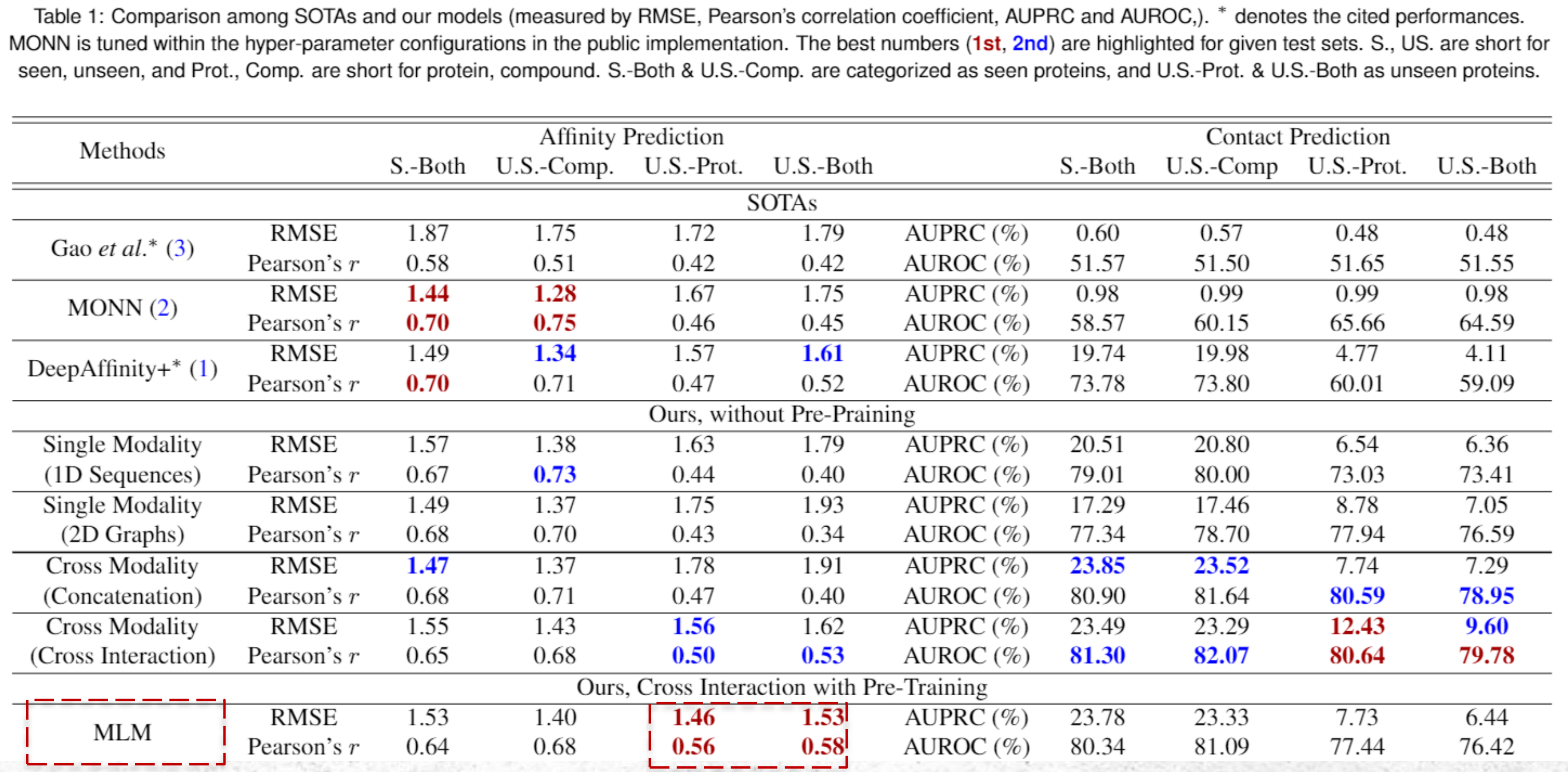


Table 1: Comparison among SOTAs and our models (measured by RMSE, Pearson's correlation coefficient, AUPRC and AUROC,). * denotes the cited performances. MONN is tuned within the hyper-parameter configurations in the public implementation. The best numbers (1st, 2nd) are highlighted for given test sets. S., US. are short for seen, unseen, and Prot., Comp. are short for protein, compound. S.-Both & U.S.-Comp. are categorized as seen proteins, and U.S.-Prot. & U.S.-Both as unseen proteins.

Methods		Affinity Prediction					Contact Prediction			
		S.-Both	U.S.-Comp.	U.S.-Prot.	U.S.-Both		S.-Both	U.S.-Comp	U.S.-Prot.	U.S.-Both
SOTAs										
Gao *et al.** (3)	RMSE	1.87	1.75	1.72	1.79	AUPRC (%)	0.60	0.57	0.48	0.48
	Pearson's r	0.58	0.51	0.42	0.42	AUROC (%)	51.57	51.50	51.65	51.55
MONN (2)	RMSE	**1.44**	**1.28**	1.67	1.75	AUPRC (%)	0.98	0.99	0.99	0.98
	Pearson's r	**0.70**	**0.75**	0.46	0.45	AUROC (%)	58.57	60.15	65.66	64.59
DeepAffinity+* (1)	RMSE	1.49	**1.34**	1.57	**1.61**	AUPRC (%)	19.74	19.98	4.77	4.11
	Pearson's r	**0.70**	0.71	0.47	0.52	AUROC (%)	73.78	73.80	60.01	59.09
Ours, without Pre-Praining										
Single Modality (1D Sequences)	RMSE	1.57	1.38	1.63	1.79	AUPRC (%)	20.51	20.80	6.54	6.36
	Pearson's r	0.67	**0.73**	0.44	0.40	AUROC (%)	79.01	80.00	73.03	73.41
Single Modality (2D Graphs)	RMSE	1.49	1.37	1.75	1.93	AUPRC (%)	17.29	17.46	8.78	7.05
	Pearson's r	0.68	0.70	0.43	0.34	AUROC (%)	77.34	78.70	77.94	76.59
Cross Modality (Concatenation)	RMSE	**1.47**	1.37	1.78	1.91	AUPRC (%)	**23.85**	**23.52**	7.74	7.29
	Pearson's r	0.68	0.71	0.47	0.40	AUROC (%)	80.90	81.64	**80.59**	**78.95**
Cross Modality (Cross Interaction)	RMSE	1.55	1.43	**1.56**	1.62	AUPRC (%)	23.49	23.29	**12.43**	**9.60**
	Pearson's r	0.65	0.68	**0.50**	**0.53**	AUROC (%)	**81.30**	**82.07**	**80.64**	**79.78**
Ours, Cross Interaction with Pre-Training										
MLM	RMSE	1.53	1.40	**1.46**	**1.53**	AUPRC (%)	23.78	23.33	7.73	6.44
	Pearson's r	0.64	0.68	**0.56**	**0.58**	AUROC (%)	80.34	81.09	77.44	76.42

Experiments. Results

- Further 2D pre-training (MLM+ GraphComp) helps contact prediction
- Deteriorating affinity prediction performance

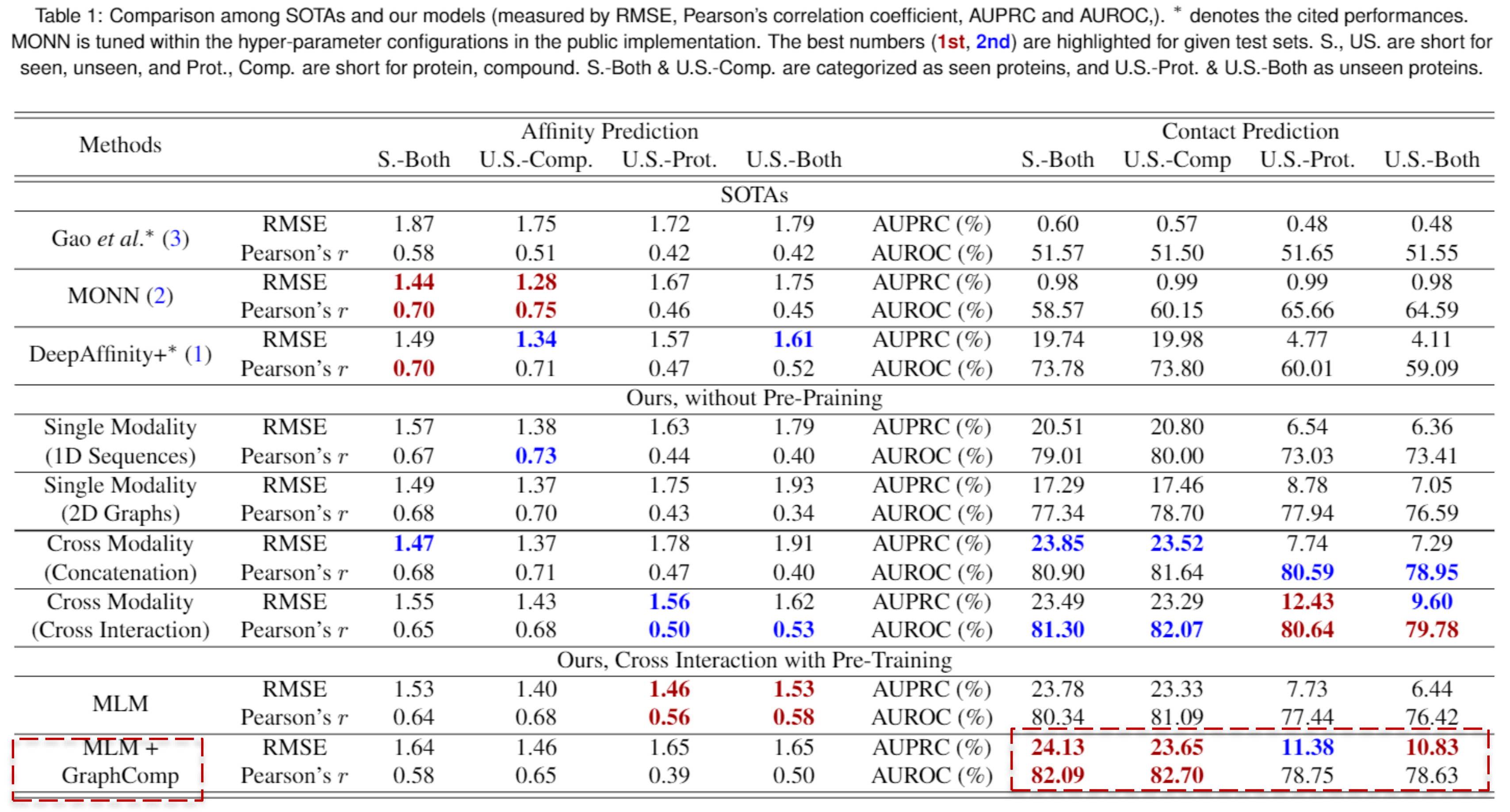


Table 1: Comparison among SOTAs and our models (measured by RMSE, Pearson's correlation coefficient, AUPRC and AUROC,). * denotes the cited performances. MONN is tuned within the hyper-parameter configurations in the public implementation. The best numbers (1st, 2nd) are highlighted for given test sets. S., US. are short for seen, unseen, and Prot., Comp. are short for protein, compound. S.-Both & U.S.-Comp. are categorized as seen proteins, and U.S.-Prot. & U.S.-Both as unseen proteins.

Methods		Affinity Prediction					Contact Prediction			
		S.-Both	U.S.-Comp.	U.S.-Prot.	U.S.-Both		S.-Both	U.S.-Comp	U.S.-Prot.	U.S.-Both
SOTAs										
Gao *et al.** (3)	RMSE	1.87	1.75	1.72	1.79	AUPRC (%)	0.60	0.57	0.48	0.48
	Pearson's r	0.58	0.51	0.42	0.42	AUROC (%)	51.57	51.50	51.65	51.55
MONN (2)	RMSE	**1.44**	**1.28**	1.67	1.75	AUPRC (%)	0.98	0.99	0.99	0.98
	Pearson's r	**0.70**	**0.75**	0.46	0.45	AUROC (%)	58.57	60.15	65.66	64.59
DeepAffinity+* (1)	RMSE	1.49	**1.34**	1.57	**1.61**	AUPRC (%)	19.74	19.98	4.77	4.11
	Pearson's r	**0.70**	0.71	0.47	0.52	AUROC (%)	73.78	73.80	60.01	59.09
Ours, without Pre-Praining										
Single Modality (1D Sequences)	RMSE	1.57	1.38	1.63	1.79	AUPRC (%)	20.51	20.80	6.54	6.36
	Pearson's r	0.67	**0.73**	0.44	0.40	AUROC (%)	79.01	80.00	73.03	73.41
Single Modality (2D Graphs)	RMSE	1.49	1.37	1.75	1.93	AUPRC (%)	17.29	17.46	8.78	7.05
	Pearson's r	0.68	0.70	0.43	0.34	AUROC (%)	77.34	78.70	77.94	76.59
Cross Modality (Concatenation)	RMSE	**1.47**	1.37	1.78	1.91	AUPRC (%)	**23.85**	**23.52**	7.74	7.29
	Pearson's r	0.68	0.71	0.47	0.40	AUROC (%)	80.90	81.64	**80.59**	**78.95**
Cross Modality (Cross Interaction)	RMSE	1.55	1.43	**1.56**	1.62	AUPRC (%)	23.49	23.29	**12.43**	**9.60**
	Pearson's r	0.65	0.68	**0.50**	**0.53**	AUROC (%)	**81.30**	**82.07**	**80.64**	**79.78**
Ours, Cross Interaction with Pre-Training										
MLM	RMSE	1.53	1.40	**1.46**	**1.53**	AUPRC (%)	23.78	23.33	7.73	6.44
	Pearson's r	0.64	0.68	**0.56**	**0.58**	AUROC (%)	80.34	81.09	77.44	76.42
MLM + GraphComp	RMSE	1.64	1.46	1.65	1.65	AUPRC (%)	**24.13**	**23.65**	**11.38**	**10.83**
	Pearson's r	0.58	0.65	0.39	0.50	AUROC (%)	**82.09**	**82.70**	78.75	78.63

Takeaways

- For inadequate data information:
  - Different modality information benefits different tasks
  - Incorporate both (cross-modality) achieves SOTA
- For insufficient supervision:
  - Different modality pre-training boosts with trade-off
  - MLM benefits affinity prediction and further +GraphComp contact

Further Discussions

- Potentials of cross-modality learning:
  - More modalities data (e.g. 3D coordinates)
  - More variants of one modality (e.g. atom graphs)

- Potentials of self-supervised learning:
  - Different pre-training strategies
  - More self-supervised labels
  - Self-supervision for more modalities

References

[1] Explainable Deep Relational Networks for Predicting Compound-Protein Affinities and Contacts

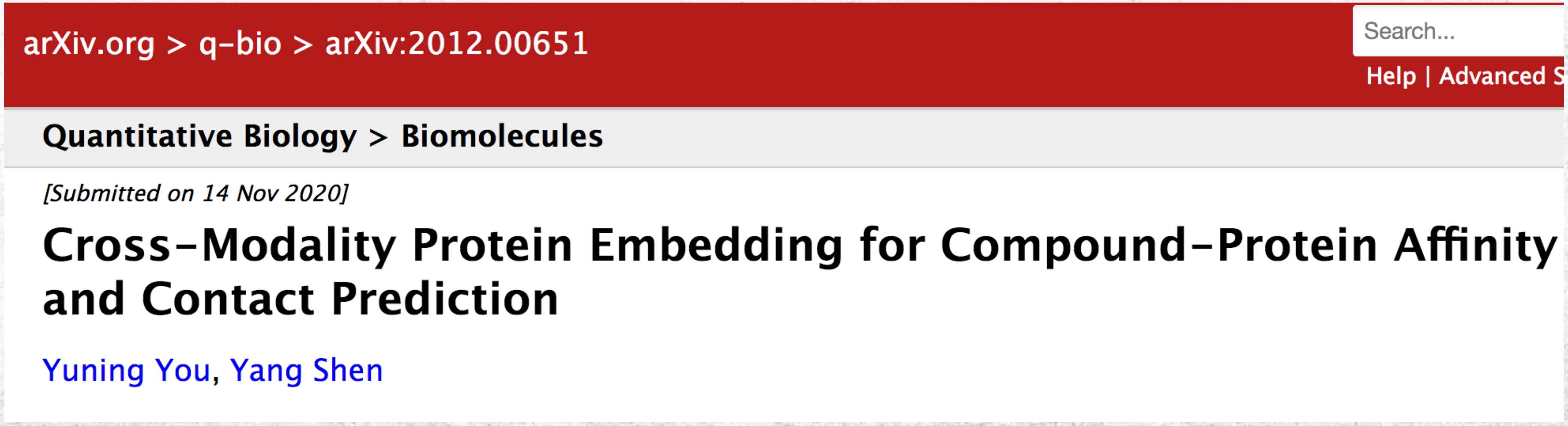
[2] RaptorX-Property: A Web Server for Protein Structure Property Prediction

[3] PDBSum: Summaries and Analyses of PDB Structures

[4] BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities

[5] The Pfam Protein Families Database

[6]



arXiv.org > q-bio > arXiv:2012.00651

Quantitative Biology > Biomolecules

[Submitted on 14 Nov 2020]

Cross-Modality Protein Embedding for Compound-Protein Affinity and Contact Prediction

Yuning You, Yang Shen

Acknowledgement

NSF

NATIONAL INSTITUTES OF HEALTH

TEXAS A&M UNIVERSITY

Ph.D. and postdoc openings

https://shen-lab.github.io

https://github.com/Shen-Lab/DeepAffinity

TEXAS A&M UNIVERSITY
Engineering

Thank you for listening!