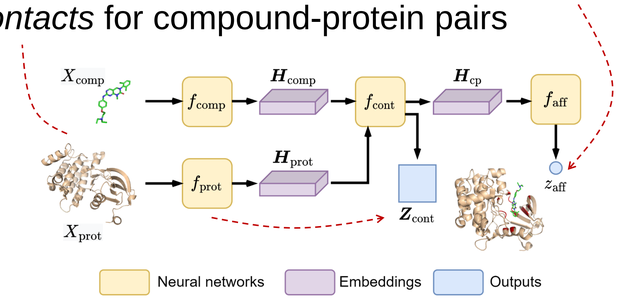
Cross-Modality and Self-Supervised Protein Embedding for Compound-Protein Affinity and Contact Prediction [1]

Yuning You, Yang Shen

TEXAS A&M UNIVERSITY

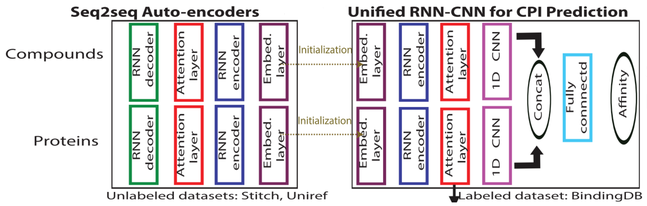
Task: Simultaneously predicting *affinities* and *contacts* for compound-protein pairs



- Most FDA-approved drug-target pairs are between molecules & proteins
- Atomic contacts (binding pockets) are interpretability for affinity prediction

Challenges: Unaware of structures, limited in labels

- Contact prediction is highly structure-dependent
- Traditional structure-free methods rely on sequence inputs

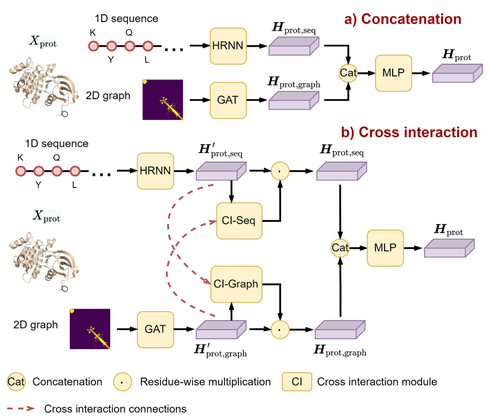


Sequence-based DeepAffinity+ [2]

- Affinity measurements for pairwise compound-protein data are sparse
- Even sparser in non-bonded atomic contacts from co-crystal structures

Solution I: Multi-modal learning to incorporate structure information

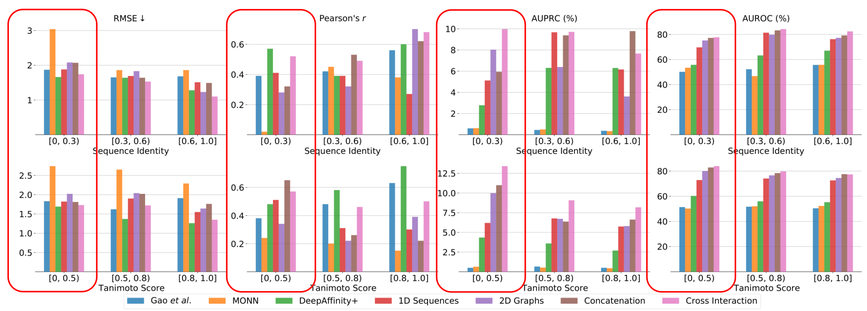
- 1D input: protein sequence
- 2D input: Protein intra-contact map
- Models:
  a) Concatenation;
  b) Cross interaction



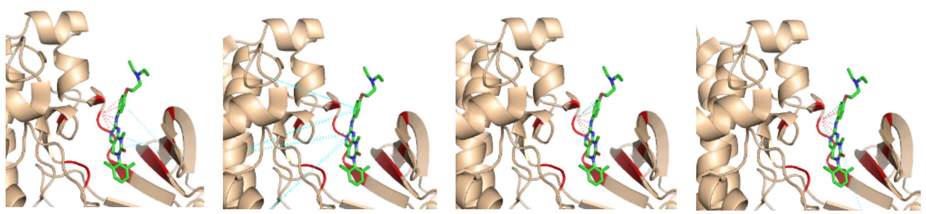
Result I: 2D modality benefits contact prediction and 1D benefits affinity. 1D+2D achieves the best

Methods	Seen-Protein Sets		Unseen-Protein Sets	
	Seen-Both	Unseen-Compound	Unseen-Protein	Unseen-Both
Affinity Prediction in RMSE (Pearson's r in parentheses)				
Gao *et al.**	1.87 (0.58)	1.75 (0.51)	1.72 (0.42)	1.79 (0.42)
MONN	**1.44** (**0.70**)	**1.28** (**0.75**)	1.67 (0.46)	1.75 (0.45)
DeepAffinity+*	1.49 (0.70)	1.34 (0.71)	1.57 (0.47)	**1.61** (0.52)
1D Sequences	1.57 (0.67)	1.38 (0.73)	1.63 (0.44)	1.79 (0.40)
Pred. 2D Graphs	1.49 (0.68)	1.37 (0.70)	1.75 (0.43)	1.93 (0.34)
True 2D Graphs	1.69 (0.59)	1.62 (0.58)	1.88 (0.33)	1.99 (0.25)
Concatenation	1.47 (0.68)	1.37 (0.71)	1.78 (0.47)	1.91 (0.40)
Cross Interaction	1.55 (0.65)	1.43 (0.68)	**1.56** (**0.50**)	1.62 (**0.53**)
Contact Prediction in AUPRC (AUROC in parentheses, %)				
Gao *et al.**	0.60 (51.57)	0.57 (51.50)	0.48 (51.60)	0.48 (51.55)
MONN	0.98 (58.57)	0.99 (60.15)	0.99 (65.66)	0.98 (64.59)
DeepAffinity+*	19.74 (73.78)	19.98 (73.80)	4.77 (60.01)	4.11 (59.09)
1D Sequences	20.51 (79.01)	20.80 (80.00)	6.54 (73.03)	6.36 (73.41)
Pred. 2D Graphs	17.29 (77.34)	17.46 (78.70)	8.78 (77.94)	7.05 (76.59)
True 2D Graphs	21.41 (**84.60**)	21.33 (**85.17**)	10.52 (**84.08**)	9.40 (**84.29**)
Concatenation	**23.85** (80.90)	**23.52** (81.64)	7.74 (80.59)	7.29 (78.95)
Cross Interaction	23.49 (81.30)	23.29 (82.07)	**12.43** (80.64)	**9.60** (79.78)

Performance advantage is preserved in the out-of-distribution domain (measured by designated similarity metrics between training and test data, see x-axis below)

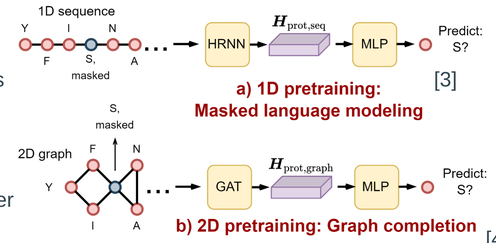


Case study for the compound-protein pair of LHL-LCK.
Cross-modality models identify contacts with higher precision



Solution II: Self-supervision for label scarcity

- Self-supervised strategies for individual modalities
- Joint pretraining conducts MLM & GraphComp together



[3] [4]

- Pretraining dataset: Pfam-A RP15 [5]
  - Smaller set (S): 60,137 sequences with measured structures
  - Larger set (L): 12,798,671 sequences

Result II: Multi-modal joint pretraining could further synergize 1D and 2D modalities

Cross Interaction	Seen-Protein Sets		Unseen-Protein Sets	
	Seen-Both	Unseen-Compound	Unseen-Protein	Unseen-Both
Affinity Prediction in RMSE (Pearson's r in parentheses)				
Non Pre-Train	1.57 (0.66)	1.46 (0.68)	1.63 (0.49)	1.64 (0.54)
MLM-S	**1.53** (**0.64**)	**1.40** (**0.68**)	**1.46** (**0.56**)	**1.53** (**0.58**)
GraphComp-S	1.62 (0.59)	1.44 (0.66)	1.60 (0.43)	1.67 (0.47)
MLM+GraphComp-S	1.64 (0.58)	1.46 (0.65)	1.65 (0.39)	1.65 (0.50)
MLM-L	1.59 (0.62)	1.46 (0.65)	1.62 (0.47)	1.63 (0.57)
MLM+GraphComp-L	1.58 (0.62)	1.45 (0.66)	1.74 (0.33)	1.85 (0.32)
Contact Prediction in AUPRC (AUROC in parentheses, %)				
Non Pre-Train	23.91 (79.48)	23.06 (80.60)	11.40 (77.73)	8.41 (76.42)
MLM-S	23.78 (80.34)	23.33 (81.09)	7.73 (77.44)	6.44 (76.42)
GraphComp-S	23.63 (79.71)	23.41 (81.31)	11.36 (76.67)	9.36 (76.00)
MLM+GraphComp-S	**24.13** (**82.09**)	**23.65** (**82.70**)	11.38 (78.75)	10.83 (78.63)
MLM-L	23.30 (80.40)	23.05 (81.18)	11.35 (81.01)	9.40 (79.46)
MLM+GraphComp-L	23.71 (81.21)	23.22 (82.33)	**13.47** (**82.00**)	**11.17** (**80.10**)

(Data not shown: Pretraining compound graphs further helped unseen sets)

More advanced pretraining techniques (e.g. GraphCL [6] for 2D modality) need to be further tailored for protein modeling

Tasks		GraphComp		GraphCL			GraphComp		GraphCL	
		S.-Both	U.S.-Comp.	S.-Both	U.S.-Comp.		U.S.-Prot.	U.S.-Both	U.S.-Prot.	U.S.-Both
Affinity prediction	RMSE	1.62	1.44	1.53	1.41	RMSE	1.60	1.67	1.66	1.77
	Pearson's r	0.59	0.66	0.67	0.70	Pearson's r	0.43	0.47	0.44	0.46
Contact prediction	AUPRC (%)	23.63	23.41	18.15	17.30	AUPRC	11.36	9.36	11.09	8.55
	AUROC (%)	79.71	81.31	76.04	75.96	AUROC	76.67	76.00	73.46	70.94

References

[1] Cross-Modality and Self-Supervised Protein Embedding for Compound–Protein Affinity and Contact Prediction, *Bioinformatics* 2022
[2] Explainable Deep Relational Networks for Predicting Compound–Protein Affinities and Contacts, *JCIM* 2020
[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *ACL* 2019
[4] When Does Self-Supervision Help Graph Convolutional Networks? *ICML* 2020
[5] The Pfam protein families database, *NAR* 2012
[6] Graph Contrastive Learning with Augmentations, *NeurIPS* 2020