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

Yuning You, Yang Shen
Texas A&M University

NEURAL INFORMATION PROCESSING SYSTEMS
Machine Learning for Structural Biology Workshop

## ➢ Background

- ❖ Computational prediction of compound-protein interactions (CPI) has been of great interest partly due to its potential impact on accelerating drug discovery, with recent progress including:
- ❖ We focus on **interpretable compound-protein affinity and contact (CPAC) prediction** without the need of compound-protein co-crystal or docked structures, even where unbound structures of proteins are not assumed here.
- ❖ We note that earlier works for this task represent proteins as **1D amino-acid sequences**, whereas structure-aware representations of proteins (such as **sequence-predicted residue-residue 2D contact maps**) can be useful.

## ➢ Contributions

- ❖ We treat protein data as available in both modalities of **1D sequences** and **(sequence-predicted) 2D contact maps**, with the following two questions asked and addressed:

  (Q1) How do the two modalities **compare** with each other for the task of structure-free interpretable CPI prediction, i.e., compound-protein affinity and contact (CPAC) prediction?

  (A1) The 1D or 2D modality of proteins **did not dominate** each other for proteins seen in the training set; however, the 1D and 2D modality-based models tend to **generalize better** for unseen proteins in affinity prediction and contact prediction, respectively

  (Q2) Is there an advantage to **exploit both modalities**?

  (A2) For the first time, we propose **cross-modality learning models** for the task of structure-free interpretable CPI prediction, to capture and fuse the different information from both 1D&2D modalities of proteins.

## ➢ Pipeline Overview

- ❖ Given a compound-protein pair, a CPAC model is targeted at making prediction for both the intermolecular affinity and (atom-residue) contacts, comprising of the following three major components:

  (1) Neural-network encoders that separately extract embeddings for the compound and protein. GNN is adopted for compound 2D chemical graphs and HRNN is chosen for protein 1D amino-acid sequences.

  (2) Interaction module taking the encoded embeddings as inputs, employing joint attention to output the interaction matrix and joint embedding to extract embeddings for compound-protein pairs.

  (3) Affinity module that predicts the affinity given the joint embedding, consisting of 1D convolutional, pooling layers, and MLP.

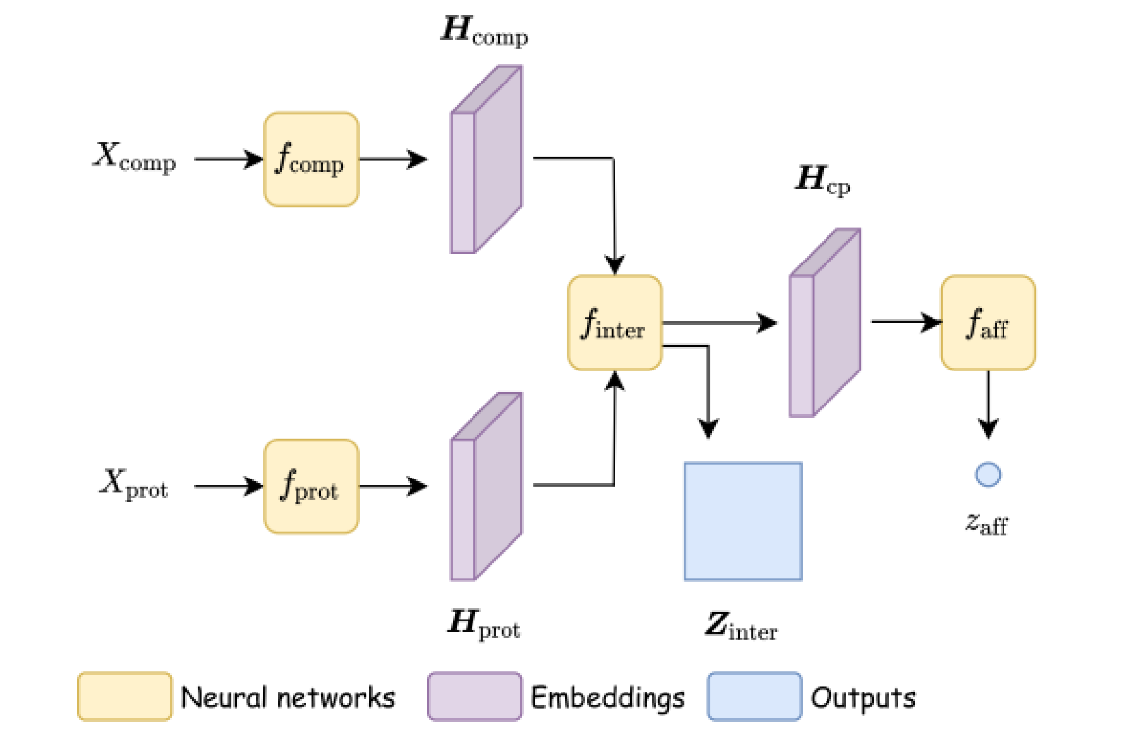

**Figure 1:** Pipeline overview for compound-protein affinity and contact prediction model $f_{CPAC}$.

- ❖ After the CPAC model forwardly generates the outputs, true labels are compared to calculate the loss. The model is trained end to end while the training loss is minimized.

## ➢ Single-Modality Models and Performances Protein

- ❖ We follow **DeepAffinity+** [1] to use **HRNN** to encode 1D amino-acid sequences, and employ an expressive GNN model, **GAT**, to process 2D contact maps.
- ❖ 2D contact map prediction is done by **RaptorX-contact** [2] that exploits both sequence and evolutionary information.
- ❖ We compare the empirical results between taking 1D amino-acid sequences and 2D contact maps as protein inputs, with the following **observations**:

  (1) For **affinity prediction**, 1D sequences and 2D graphs did not yield major differences especially in Pearson's r. One conjecture is that affinity prediction for unseen-protein cases are **not as challenging as** intermolecular contact prediction to show the benefit of the 2D modality.

  (2) For **contact prediction**, encoding proteins as 1D sequences performed better in seen proteins, (i.e. the proteins in compound-protein pairs at the inference phase are involved in the training compound-protein pairs). Meanwhile, encoding 2D protein contact maps (graphs) outperformed doing that to 1D protein sequences for unseen proteins. We conjecture that the sequential information learned from the encoder could be **more accurate** toward intermolecular contact prediction for close or even distant homologs of seen proteins but it is **less general** to unseen proteins.

**Table 1:** Affinity and contact prediction with different modalities of proteins as inputs.

| | | 1D Sequences | | 2D Graphs | |
|---|---|---|---|---|---|
| | | Test (Seen-Protein) | Unseen-Protein | Test (Seen-Protein) | Unseen-Protein |
| Affinity Prediction | RMSE ↓ | 1.57 | 1.63 | 1.49 | 1.75 |
| | Pearson's $r$ ↑ | 0.67 | 0.44 | 0.68 | 0.43 |
| Contact Prediction | AUPRC (%) ↑ | 20.51 | 6.54 | 17.29 | 8.78 |
| | AUROC (%) ↑ | 79.01 | 73.03 | 77.34 | 77.94 |

## ➢ Cross-Modality Models

- ❖ Since both sequential dependency in 1D amino-acid sequences and structural topology in 2D contact maps are important information for proteins, it is natural to propose a cross-modality learning framework that **captures and fuses the information from 1D & 2D modalities** for better performances.

  (1) Concatenation. A simple fusion model is to concatenate the extracted embeddings of the 1D and 2D modalities that are encoded by HRNN and GAT, respectively. Concatenation is commonly used in previous work to **preserve information** from different sources. The concatenated output is fed to a MLP for the final protein embedding.

  (2) Cross interaction. Although the concatenation strategy preserves the information of individual modalities, the encoding processes for the two modalities are separate. However, the different modalities of proteins are **intrinsically correlated** with each other and could be coupled in a properly-designed representation-learning process. Therefore, we have introduced a cross interaction module to facilitate the encoder to learn protein embeddings from correlated data (1D and 2D modalities).

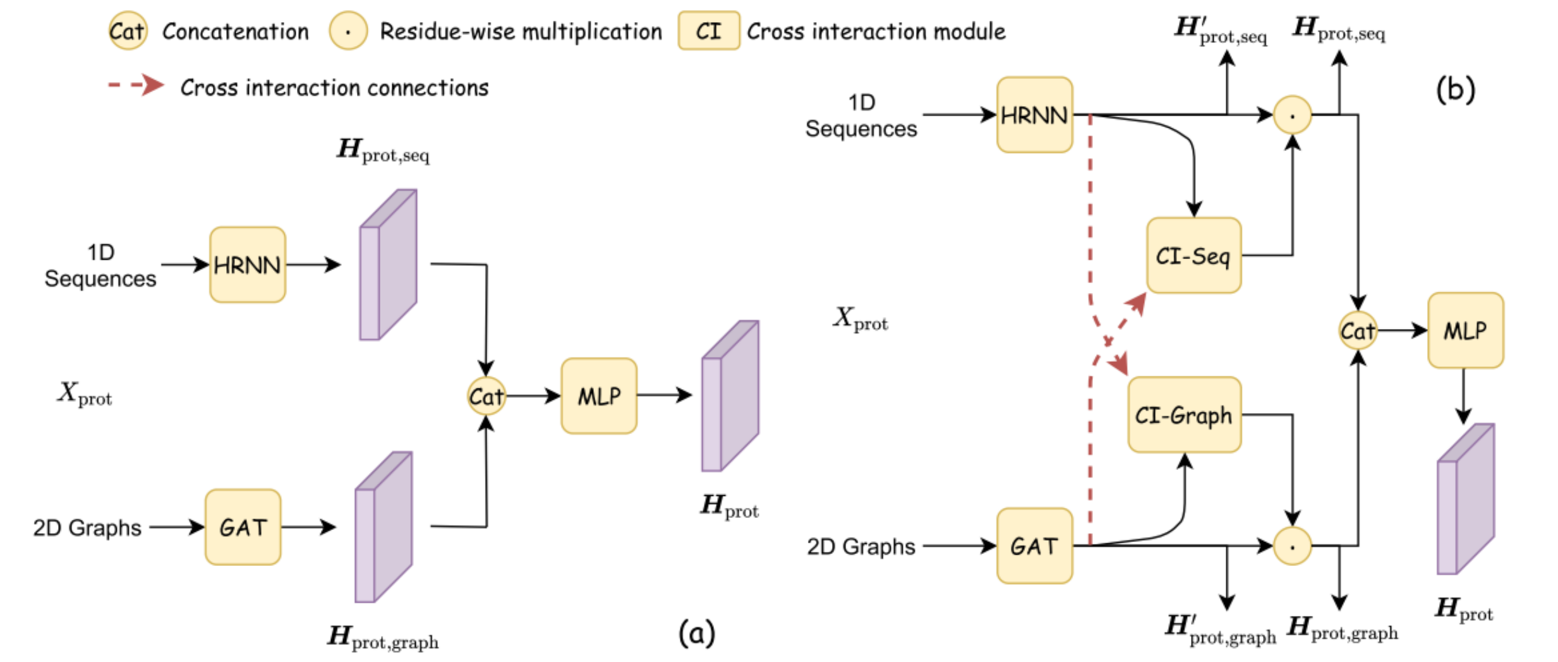

**Figure 2:** Cross-modality encoder for proteins to capture and fuse different modality information, with (a) naïve concatenation and (b) cross interaction introduced.

## ➢ Experiment Results

- ❖ We compare our single-modality and cross-modality models with two latest SOTAs for the CPAC problem, with tasks involving affinity, contact, and binding-site predictions.
- ❖ Our experiments show that cross-modality models can exploit the correlation between both modalities and enjoy the benefits of both modalities even when a simple concatenation strategy is adopted for the two embeddings.

**Table 2:** Comparison among SOTAs and our models in compound-protein affinity prediction (measured by RMSE and Pearson's correlation coefficient). * denotes the cited performances. Boldfaced were the best performances for given test sets.

| | | Test (Seen-Both) | Unseen-Compound | Unseen-Protein | Unseen-Both |
|---|---|---|---|---|---|
| SOTAs | | | | | |
| Gao et al.* | RMSE | 1.87 | 1.75 | 1.72 | 1.79 |
| | Pearson's $r$ | 0.58 | 0.51 | 0.42 | 0.42 |
| DeepAffinity+* | RMSE | 1.49 | **1.34** | 1.57 | 1.61 |
| | Pearson's $r$ | **0.70** | 0.71 | 0.47 | 0.52 |
| Ours | | | | | |
| Single Modality (1D Sequences) | RMSE | 1.57 | 1.38 | 1.63 | 1.79 |
| | Pearson's $r$ | 0.67 | **0.73** | 0.44 | 0.402 |
| Single Modality (Pred. 2D Graphs) | RMSE | 1.49 | 1.37 | 1.75 | 1.93 |
| | Pearson's $r$ | 0.68 | 0.70 | 0.43 | 0.34 |
| Single Modality (True 2D Graphs) | RMSE | 1.69 | 1.62 | 1.88 | 1.99 |
| | Pearson's $r$ | 0.59 | 0.58 | 0.33 | 0.25 |
| Cross Modality (Concatenation) | RMSE | **1.47** | 1.37 | 1.78 | 1.91 |
| | Pearson's $r$ | 0.68 | 0.71 | 0.47 | 0.40 |
| Cross Modality (Cross Interaction) | RMSE | 1.55 | 1.43 | **1.56** | **1.62** |
| | Pearson's $r$ | 0.65 | 0.68 | **0.50** | **0.53** |

**Table 3:** Comparison among SOTAs and our models in contact prediction (measured by AUPRC and AUROC). * denotes the cited performances. Boldfaced were the best performances for given test sets.

| | | Test (Seen-Both) | Unseen-Compound | Unseen-Protein | Unseen-Both |
|---|---|---|---|---|---|
| SOTAs | | | | | |
| Gao et al.* | AUPRC (%) | 0.60 | 0.57 | 0.48 | 0.48 |
| | AUROC (%) | 51.57 | 51.50 | 51.65 | 51.55 |
| DeepAffinity+* | AUPRC (%) | 19.74 | 19.98 | 4.77 | 4.11 |
| | AUROC (%) | 73.78 | 73.80 | 60.01 | 59.09 |
| Ours | | | | | |
| Single Modality (1D Sequences) | AUPRC (%) | 20.51 | 20.80 | 6.54 | 6.36 |
| | AUROC (%) | 79.01 | 80.00 | 73.03 | 73.41 |
| Single Modality (Pred. 2D Graphs) | AUPRC (%) | 17.29 | 17.46 | 8.78 | 7.05 |
| | AUROC (%) | 77.34 | 78.70 | 77.94 | 76.59 |
| Single Modality (True 2D Graphs) | AUPRC (%) | 21.41 | 21.33 | 10.52 | 9.40 |
| | AUROC (%) | **84.60** | **85.17** | **84.08** | **84.29** |
| Cross Modality (Concatenation) | AUPRC (%) | **23.85** | **23.52** | 7.74 | 7.29 |
| | AUROC (%) | 80.90 | 81.64 | 80.59 | 78.95 |
| Cross Modality (Cross Interaction) | AUPRC (%) | 23.49 | 23.29 | **12.43** | **9.60** |
| | AUROC (%) | 81.30 | 82.07 | 80.64 | 79.78 |

**Table 4:** Comparison among SOTAs and our models in ligand-specific and structure-free protein binding-site prediction. * denotes the cited numbers. Boldfaced are the best performances for individual test sets.

| | | Test (Seen-Both) | Unseen-Compound | Unseen-Protein | Unseen-Both |
|---|---|---|---|---|---|
| SOTAs | | | | | |
| Gao et al.* | AUPRC (%) | 5.43 | 5.38 | 4.95 | 4.96 |
| | AUROC (%) | 49.79 | 50.51 | 48.21 | 48.74 |
| DeepAffinity+* | AUPRC (%) | 42.16 | 43.14 | 16.98 | 15.65 |
| | AUROC (%) | 76.33 | 78.22 | 64.93 | 65.18 |
| Ours | | | | | |
| Single Modality (1D Sequences) | AUPRC (%) | 40.35 | 40.81 | 20.37 | 20.17 |
| | AUROC (%) | 76.69 | 77.79 | 70.28 | 70.96 |
| Single Modality (Pred. 2D Graphs) | AUPRC (%) | 33.17 | 33.83 | 25.57 | 22.49 |
| | AUROC (%) | 75.11 | 76.53 | 76.15 | 74.87 |
| Single Modality (True 2D Graphs) | AUPRC (%) | 41.73 | 42.58 | 29.44 | **29.02** |
| | AUROC (%) | **83.67** | **84.85** | **83.82** | **84.15** |
| Cross Modality (Concatenation) | AUPRC (%) | **43.56** | **44.12** | 28.15 | 26.44 |
| | AUROC (%) | 78.83 | 79.75 | 78.51 | 77.61 |
| Cross Modality (Cross Interaction) | AUPRC (%) | 43.45 | 43.00 | **30.54** | 27.18 |
| | AUROC (%) | 78.85 | 79.73 | 77.37 | 77.54 |

## ➢ References

[1] Mostafa Karimi, Di Wu, Zhangyang Wang, Yang Shen. "Explainable Deep Relational Networks for Predicting Compound-Protein Affinities and Contacts", arXiv 2019.
[2] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, Jinbo Xu. "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model", PLOS Computational Biology 2017.

Contact: {yuning.you,yshen}@tamu.edu