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

Yuning You, Yang Shen
Texas A&M University
{yuning.you,yshen}@tamu.edu

Abstract

Compound-protein pairs dominate FDA-approved drug-target pairs and the prediction of compound-protein affinity and contact (CPAC) could help accelerate drug discovery. In this study we consider proteins as multi-modal data including 1D amino-acid sequences and (sequence-predicted) 2D residue-pair contact maps. We empirically evaluate the embeddings of the two single modalities in their accuracy and generalizability of CPAC prediction (i.e. structure-free interpretable compound-protein affinity prediction). And we rationalize their performances in both challenges of embedding individual modalities and learning generalizable embedding-label relationship. We further propose two models involving cross-modality protein embedding and establish that the one with cross interaction (thus capturing correlations among modalities) outperforms SOTAs and our single modality models in affinity, contact, and binding-site predictions for proteins never seen in the training set.

1 Introduction

Computational prediction of compound-protein interactions (CPI) has been of great interest partly due to its potential impact on accelerating drug discovery [1, 2]. Recent progress in this topic includes (1) the improved accuracy of structure-based binary classification [3, 4] and affinity regression [5, 6] for CPI; (2) the structure-free inputs that remove the demand of compound-protein co-crystal or docked structures that are experimentally or computationally expensive [7, 8, 9, 10, 11]; and (3) the recent development of interpretable structure-free predictions of both protein-ligand binding affinities and their atomic contacts [9, 12, 13].

We focus on interpretable CPI prediction without the need of compound-protein co-crystal or docked structures. Even unbound structures of proteins are not assumed here. Specifically, we aim at simultaneous prediction of compound-protein affinity and contacts in the aforementioned structure-free setting. We note that earlier works for this task represent proteins as 1D amino-acid sequences [12, 13] or 1D structurally-annotated sequences [9]. However, 1D sequences of proteins adopt 3D structures to function, including interactions with compounds; so structure-aware representations of proteins (such as sequence-predicted residue-residue 2D contact maps) can also be useful, as explored in a recent affinity predictor [11]. (Although compound data can be available in both modalities of 1D SMILES and chemical graphs, we did not pursue both modalities and only represented compounds as graphs because SMILES strings have limited descriptive power and known worse performance in the CPAC task [9, 12].)

In this paper, we treat protein data as available in both modalities of 1D sequences and (sequence-predicted) 2D contact maps. And we ask the following questions: 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? Is there an advantage to exploit both modalities? And what would be a beneficial cross-modality approach? Our contributions and findings include the following:

Machine Learning for Structural Biology Workshop, NeurIPS 2020, Vancouver, Canada.

- By embedding either modality with recurrent or graph neural networks and predicting affinities through intermolecular contact-predicting joint attentions, we empirically compared the two resulting single-modality models and found that: 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. We further provided conjectures involving the difficulty of embedding various modality and the mappings between various embeddings and affinity or contact labels.

- 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. And we empirically demonstrate that the two cross-modality learning models (through concatenation or cross-interaction of sequence and graph embeddings) achieve better accuracy and generalizability compared to the state of the art (SOTA) and our single-modality models, in compound-protein affinity, contact, and binding-site prediction.

2 Pipeline Overview

We assume that compounds are available in (1D SMILES or) 2D chemical graphs and proteins available in 1D amino-acid sequences. Given a compound-protein pair $(X_{\text{comp}}, X_{\text{prot}})$ composed of N_{comp} atoms and N_{prot} residues where $N_{\text{comp}}, N_{\text{prot}}$ are predefined and fixed numbers (padding is applied to ensure the fixed sizes), a CPAC model $f_{\text{CPAC}} : \mathbb{X}_{\text{comp}} \times \mathbb{X}_{\text{prot}} \to \mathbb{R}_{\geq 0} \times [0,1]^{N_{\text{comp}} \times N_{\text{prot}}}$ is targeted at making prediction for both the intermolecular affinity z_{aff} and (atom-residue) contacts $\boldsymbol{Z}_{\text{inter}}$, where $\mathbb{X}_{\text{comp}}, \mathbb{X}_{\text{prot}}$ are respectively the spaces for $X_{\text{comp}}, X_{\text{prot}}$. The SOTA pipelines for CPAC [12, 9, 13] comprise of the following three major components as shown in Figure 1.

(1) **Neural-network encoders** $f_{\text{comp}} : \mathbb{X}_{\text{comp}} \to \mathbb{R}^{N_{\text{comp}} \times D}, f_{\text{prot}} : \mathbb{X}_{\text{prot}} \to \mathbb{R}^{N_{\text{prot}} \times D}$ that separately extract embeddings $\boldsymbol{H}_{\text{comp}}, \boldsymbol{H}_{\text{prot}}$ for the compound X_{comp} and protein X_{prot} where D is hidden dimension. Graph neural network (GNN, [14, 15, 16, 17, 18, 19]) is adopted for compound 2D chemical graphs and hierarchical recurrent neural network (HRNN, [20]) is chosen for protein 1D amino-acid sequences.

(2) **Interaction module** $f_{\text{inter}} : \mathbb{R}^{N_{\text{comp}} \times D} \times \mathbb{R}^{N_{\text{prot}} \times D} \to [0,1]^{N_{\text{comp}} \times N_{\text{prot}}} \times \mathbb{R}^{L \times D}$ taking the encoded embeddings $\boldsymbol{H}_{\text{comp}}, \boldsymbol{H}_{\text{prot}}$ as inputs, employing joint attention to output the interaction matrix $\boldsymbol{Z}_{\text{inter}}$ and joint embedding to extract embeddings $\boldsymbol{H}_{\text{cp}}$ for compound-protein pairs, where L is hidden length determined by $N_{\text{comp}}, N_{\text{prot}}$.

(3) **Affinity module** $f_{\text{aff}} : \mathbb{R}^{L \times D} \to \mathbb{R}$ that predicts the affinity z_{aff} given the joint embedding $\boldsymbol{H}_{\text{cp}}$, consisting of 1D convolutional, pooling layers, and multi-layer perceptron (MLP). Note that the contact-predicting interaction module feeds the affinity module, making affinity prediction intrinsically interpretable by the underlying contacts.

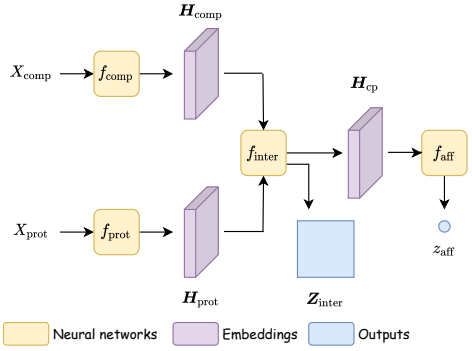

Figure 1: Pipeline overview for compound-protein affinity and contact prediction model f_{CPAC}.

After the CPAC model f_{CPAC} forwardly generates the outputs $(z_{\text{aff}}, \boldsymbol{Z}_{\text{inter}})$, true labels $(y_{\text{aff}}, \boldsymbol{Y}_{\text{inter}})$ are compared to calculate the loss, l_{CPAC}, which consists of affinity loss l_{aff}, intermolecular atom–residue contact/interaction loss l_{inter} and three structure-aware sparsity regularization loss $l_{\text{group}}, l_{\text{fused}}, l_{\text{L1}}$ described in [12], expressed as:

$$l_{\text{CPAC}} = l_{\text{aff}} + \lambda_{\text{inter}} l_{\text{inter}} + \lambda_{\text{group}} l_{\text{group}} + \lambda_{\text{fused}} l_{\text{fused}} + \lambda_{\text{L1}} l_{\text{L1}}. \tag{1}$$

The model is trained end to end while the training loss is minimized. More details can be found in [12].

3 Single-Modality Models and Performances

Protein 1D sequences. We follow DeepAffinity+ [12] as described above and use HRNN to encode protein sequences. One change we made was replacing the hierarchical joint attention with naïve joint attention in the interaction module expressed as:

$$\boldsymbol{Z}_{\text{inter}} = \boldsymbol{Z}'_{\text{inter}}/\text{sum}(\boldsymbol{Z}'_{\text{inter}}), \quad z'_{\text{inter},i,j} = (\boldsymbol{h}_{\text{comp},i}\boldsymbol{W}_{\text{comp,attn}})^{\text{T}}(\boldsymbol{h}_{\text{prot},j}\boldsymbol{W}_{\text{prot,attn}}), \tag{2}$$

where $z_{i,j} = \boldsymbol{Z}[i,j], \boldsymbol{h}_i = \boldsymbol{H}[i,:], i = 1,...,N_{\text{comp}}, j = 1,...,N_{\text{prot}}$ and $\boldsymbol{W}_{\text{comp,attn}}, \boldsymbol{W}_{\text{prot,attn}}$ are two learnable attention matrices.

Protein 2D contact maps. In previous SOTAs for CPAC, proteins are often represented as 1D amino-acid sequences [12, 9, 13]. We propose to adopt the 2D modality of proteins as inputs and model them as graphs with the following reasons. Firstly, graphs are structure-aware compared with 1D sequences, potentially resulting in better generalizability. Secondly, graphs are more concise yet informative (focusing on pairwise residue interactions) compared to the data structure of 3D coordinates (which are also harder to predict than contact maps) [21]. Lastly, the recent surge of development in graph learning [14, 15, 16] provides advanced tools to facilitate graph representation learning.

Thus, given 2D residue-residue contact maps, we represent a protein input X_{prot} as a graph $\mathcal{G}_{\text{prot}} = \{\mathcal{V}_{\text{prot}}, \mathcal{E}_{\text{prot}}\}$ where vertices stand for residues and edges exist between residues predicted to be in contact (Z-scores of predicted probability are above 3). When actual protein contact graphs are used for comparison, the edge criteria (for residue pairs in contact) is if their C_β atoms are within 8Å. As the graphs are defined by the 2D contact maps, we may refer to them as 2D maps or 2D graphs interchangeably.

The graphs are associated with feature matrix $\boldsymbol{F}_{\text{prot}} \in \mathbb{R}^{N_{\text{prot}} \times D}$ (embedded amino-acid types of residues) and the adjacency matrix $\boldsymbol{A}_{\text{prot}} \in \{0,1\}^{N_{\text{prot}} \times N_{\text{prot}}}$ (binary contact map). We employ an expressive GNN model, graph attention network (GAT, [14]) with K layers as the protein encoder f_{prot} to extract graph embeddings, with the formulation of each layer's forward propagation as:

$$\begin{aligned} &\boldsymbol{H}^{(k)}_{\text{prot}} = \text{MLP}(\tilde{\boldsymbol{S}}^{(k-1)}\boldsymbol{H}^{(k-1)}_{\text{prot}}), \quad \tilde{\boldsymbol{S}}^{(k-1)} = {\boldsymbol{D}^{(k-1)}}^{-1}(\boldsymbol{S}^{(k-1)} \odot \boldsymbol{A}_{\text{prot}}), \\ &\boldsymbol{S}^{(k-1)} = \exp(\boldsymbol{H}^{(k-1)}_{\text{prot}}\boldsymbol{W}^{(k-1)}{\boldsymbol{H}^{(k-1)}_{\text{prot}}}^{\text{T}}), \end{aligned} \tag{3}$$

where $\boldsymbol{H}_{\text{prot}} = \boldsymbol{H}^{(K)}_{\text{prot}}, \boldsymbol{H}^{(0)}_{\text{prot}} = \boldsymbol{F}_{\text{prot}}$, the normalization matrix $\boldsymbol{D}^{(k-1)} = \text{diag}((\boldsymbol{S}^{(k-1)} \odot \boldsymbol{A}_{\text{prot}})\boldsymbol{J}_{N_{\text{prot}},1})$, $\odot$ is the element-wise multiplication, $\boldsymbol{J}_{N_{\text{prot}},1}$ is an all-ones matrix with size $N_{\text{prot}} \times 1$, and $\boldsymbol{W}^{(k-1)}$ is a learnable weight matrix.

As (unbound or ligand-bound) structure data is not readily available for many proteins, we use sequence-predicted 2D contact maps to overcome the limitation and broaden our models' applicability. 2D contact map prediction is done by RaptorX-contact [22] that exploits both sequence and evolutionary information.

Data set. We use the dataset and splitting scheme as in DeepAffinity+ [12], which is curated based on PDBbind [23] and BindingDB [24]. It contains protein sequences, predicted (and actual bound) protein contact maps, compound SMILEs and graphs, affinity labels ($\text{p}K_d$/$\text{p}K_i$) and intermolecular atomic interactions/contacts (curated from the LigPlot service of PDBsum [25]). The updated dataset is diverse: it consists of 4,446 pairs between 3,672 compounds (of wide range of properties such as logP, molecular weight, and affinity labels) and 1,287 proteins (including enzymes across all six classes, GPCRs, nuclear receptors, ion channels, and so on). The dataset is split into subsets of various challenging levels in generalizability: 795 pairs involving unseen proteins (proteins not present in the training set), 521 pairs involving unseen compounds, and 205 for unseen both; whereas the rest is randomly split into training including validation (2,334) and the default test (591) sets. Note that the default test set contains compounds or protein seen in the training set but never training compound-protein pairs.

Model training and hyperparameter tuning. We train our models end to end with the following optimization settings as in [12]: the optimizer Adam with a learning rate of 0.001, the batch size of 64 and the maximum amount of training epochs being 200. The best checkpoint model is selected via validation. The following hyperparameters in the loss function are optimized following a two-stage process over pre-defined grids [12]. Specifically, λ_{group}, λ_{fused}, and λ_{L1} are first tuned over

$\{0.01, 0.001, 0.0001\}$ with $\lambda_{\text{inter}} = 0$ (affinity regression alone), where the best affinity loss l_{aff} is recorded and λ_{group}, λ_{fused}, and λ_{L1} are optimized with the best AUPRC such that the corresponding affinity RMSE does not deteriorate more than 10% of the best affinity RMSE. In the second stage, we fix the optimal λ_{group}, λ_{fused}, and λ_{L1} and tune λ_{inter} over $\{1\text{e}0, 1\text{e}1, 1\text{e}2, 1\text{e}3, 1\text{e}4, 1\text{e}5\}$ based on the best AUPRC performance while jointly optimizing the regularized affinity and contact losses.

Numerical comparison of different modalities. We compare the empirical results in Table 1 between taking 1D amino-acid sequences and 2D contact maps as protein inputs, using HRNN and GAT as encoders for proteins, respectively. We make the following observations.

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

(i) For affinity prediction (see RMSE & Pearson), 1D sequences and 2D graphs did not yield major differences especially in Pearson's r. 1D sequences led to less deterioration in RMSE from the validation set (containing seen proteins) to unseen proteins.

One conjecture is that the information in graphs might be more difficult to learn compared to sequences (the training RMSE losses are 0.71 & 0.99 for 1D & 2D modalities, respectively, when long enough training processes were performed). Moreover, affinity prediction for unseen-protein cases are not as challenging as intermolecular contact prediction to show the benefit of the 2D modality (see (ii) below), as contact prediction often involves tens of thousands of values (rather than a single value) to fit for each compound-protein pair.

(ii) For contact prediction (see AUPRC & AUROC), encoding proteins as 1D sequences performed better (+3.22% at AUPRC and +1.67% at AUROC) 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 (+4.91% at AUPRC and +2.24% at AUROC) for unseen proteins.

We conjecture that sequential dependency information encoded in 1D amino-acid sequences is well captured especially for seen proteins whose embeddings are well constructed after training (as they are already represented in the training set), leading to the better contact predictions for seen proteins. However, 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.

In contrast, we conjecture that the better generalizability of the 2D modality model might result from the quality of the encoded embedding of proteins, which is co-determined by both the inputs (2D maps) and encoders (GAT models). The structural topology information encoded in protein 2D contact maps is more difficult for graph neural networks to capture even for seen proteins, leading to the worse contact predictions for seen proteins. However, such information can generalize to unseen proteins well toward contact prediction. In particular, even when sequence similarity for non-homologous proteins (to training ones) is too low to be detectable using RNNs, binding-pocket (subgraph) similarity could still preserve and be detected in 2D contact maps using GNNs thus eventually leads to better intermolecular contact prediction.

4 Cross-Modality Models

We have shown that both sequential dependency in 1D amino-acid sequences and structural topology in 2D contact maps are important information for proteins to extract accurate and generalizable embeddings. Therefore it is natural to propose a cross-modality learning framework that captures and fuses the information from 1D & 2D modalities for better performances. Specifically we have designed the following two models.

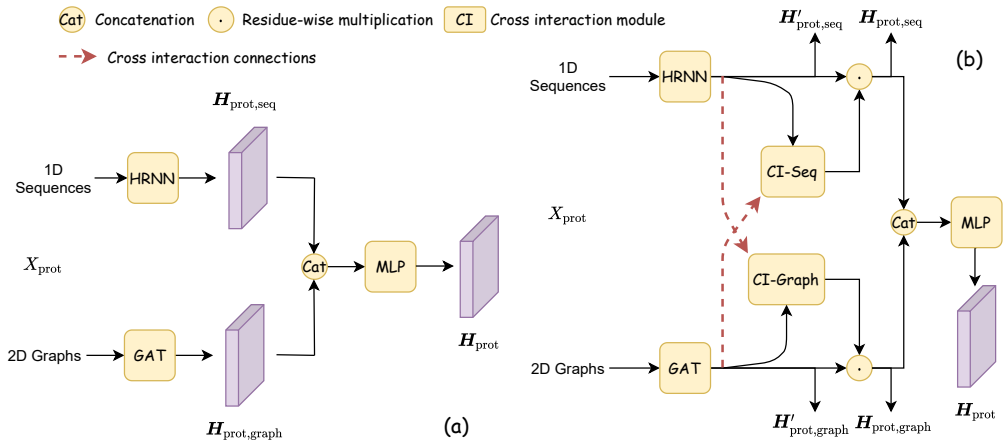

Figure 2: Cross-modality encoder for proteins to capture and fuse different modality information, with (a) naïve concatenation and (b) cross interaction introduced.

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, as shown in Figure 2(a). Indeed, concatenation is commonly used in previous work [26, 27] to preserve information from different sources. The concatenated output is fed to a multi-layer perception (MLP) for the final protein embedding $\boldsymbol{H}_{\text{prot}}$.

Cross interaction. Although the aforementioned concatenation strategy preserves the information of individual modalities, the encoding processes for the two modalities are separate. In other words, the two types of embeddings from different modalities were independently encoded and then mixed through concatenation. 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), as shown in Figure 2(b). Specifically, given the outputs of encoders $\boldsymbol{H}'_{\text{prot,seq}}$ and $\boldsymbol{H}'_{\text{prot,graph}}$, we calculate sequence & graph cross-modality outputs $\boldsymbol{H}_{\text{prot,seq}}$ and $\boldsymbol{H}_{\text{prot,graph}}$, respectively:

$$\boldsymbol{h}_{\text{prot,seq},n} = (\text{sigmoid}(\boldsymbol{h}''_{\text{prot,graph},n}{}^{\text{T}} \boldsymbol{h}'_{\text{prot,seq},n}) + 1)\boldsymbol{h}'_{\text{prot,seq},n}, \tag{4}$$

$$\boldsymbol{h}_{\text{prot,graph},n} = (\text{sigmoid}(\boldsymbol{h}''_{\text{prot,seq},n}{}^{\text{T}} \boldsymbol{h}'_{\text{prot,graph},n}) + 1)\boldsymbol{h}'_{\text{prot,seq},n}, \tag{5}$$

where $\boldsymbol{h}^{\cdot}_{n} = \boldsymbol{H}^{\cdot}[n,:]$ ($\cdot$ can be empty, $'$ or $''$), $\boldsymbol{H}''_{\text{prot,graph}} = \boldsymbol{H}'_{\text{prot,graph}}\boldsymbol{W}_{\text{cross,graph}}$, $\boldsymbol{H}''_{\text{prot,seq}} = \boldsymbol{H}'_{\text{prot,seq}}\boldsymbol{W}_{\text{cross,seq}}$, and $\boldsymbol{W}_{\text{cross,seq}}$ and $\boldsymbol{W}_{\text{cross,graph}}$ are learnable weight matrices. Instead of independently extracting information from protein modalities (1D sequences and 2D contact maps), the cross interaction module enforces a learned relationship between the encoded embeddings of the two protein modalities, which is expected to better capture the information from the correlated protein modalities and to benefit the affinity and contact prediction. Again, $\boldsymbol{H}_{\text{prot,seq}}$ and $\boldsymbol{H}_{\text{prot,graph}}$ (now with information from each other) are concatenated and fed to an MLP for the final protein embedding $\boldsymbol{H}_{\text{prot}}$.

The idea of cross interaction was previously introduced in [28] and modified in our study as follows. First, we do not normalize cross interaction along residues (sequence length is 1000 here) since it would significantly change the scale of the residue embeddings. Second, we restrict the cross interaction for each residue in the range of [0, 1] with sigmoid function to represent the cross-modality "interaction strength".

5 Results

We compare our single-modality and cross-modality models with two latest SOTAs for the CPAC problem, namely Gao et al. [8] and DeepAffinity+ [12]. Tasks involved include affinity, contact, and binding-site predictions.

Affinity and Contact Prediction. As shown in Table 2 and 3, compared to SOTAs, our models have achieved similar performances in affinity prediction (RMSE and Pearson's r) and improved performances in contact prediction (AUPRC and AUROC) especially for proteins never seen in training (unseen-protein and unseen-both). We have made the following observations.

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

First, our models used similar backbone as DeepAffinity+ and revised the joint attention mechanism; thus DeepAffinity+ and our 1D sequence-based single-modality model, both using protein sequences, had similar performances in affinity prediction but ours improved contact prediction.

Second, as observed in Section 3, compared to the 1D modality of protein sequences, the 2D modality of (sequence-predicted) protein contact maps improved the generalizability of compound-protein contact prediction for unseen proteins or unseen both, even though it resulted in slightly worse accuracy for seen proteins. Higher-quality actual protein contact maps, compared to sequence-predicted ones, further benefited contact prediction for both seen and unseen proteins; but they could lead to worse affinity prediction. These results echo our earlier conjecture that structural topology in the 2D graphs is more informative for the more complex task of contact prediction even though it may not be as effective as the 1D sequences for the less complex task of affinity prediction.

We have also made the following observations for our cross-modality fusion models where only sequence-predicted protein contact maps are used.

Third, fusing two modalities' information together, even by a simple concatenation strategy, could get the best of both modalities: the cross modality model by concatenation had better contact prediction

than single-modality models (even the true 2D map-based one) and a trade-off in affinity predictions (better than the 2D single modality models and worse than the 1D single modality model). These results confirm our rationale of proposing cross-modality protein encoders for the CPAC task.

Last, enforcing a learned correlation between the 1D and 2D embeddings rather than independently learning two individual embeddings, the cross-modality model with cross interaction further improved affinity prediction and actually had the best affinity accuracy among all methods for unseen proteins or unseen both. Moreover, it impressively achieved the best AUPRC for unseen proteins and unseen both. We note that, as intermolecular contacts only represent a minority (around 0.4%) of all compound-protein atom-residue pairs, AUPRC is a much more relevant measure than AUROC for contact prediction. These results reinforced our rationale that the learned correlation between embeddings from different modalities can better capture the correlated data and better perform CPAC predictions.

Protein binding-site prediction. We also compare Gao et al., DeepAffinity+, and our models for protein binding site prediction that is ligand-specific and structure-free. Our models again significantly improve the accuracy here compared to SOTAs. As actual protein structures (unbound or bound) are not assumed available, the single-modality model using true 2D contact maps (from compound-bound protein structures) here is essentially providing an estimate of the performance upper bound for unseen proteins. Impressively, using only protein sequences and sequence-predicted contact maps, both cross-modality models improved against the single modality model (true 2D graphs) for seen proteins and performed closely to the latter for unseen proteins. The cross-modality model with cross interaction achieved the best AUPRC for unseen proteins among all models compared. Again, as protein binding-site residues represent a minority among all residues, AUPRC is a much more relevant measure than AUROC for assessing binding-site prediction.

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

6 Conclusions

We explore in this study various protein modalities (1D sequences and 2D residue-residue contact maps) in the context of compound-protein affinity and contact prediction. To this end, we have exploited RNNs and GNNs to encode the 1D and 2D modalities respectively and proposed cross-modality models (concatenation and cross interaction) on top of the single-modality models.

Our experiments show that the two different protein modalities result in different accuracy and generalizability in affinity and contact predictions. Specifically, sequential dependency learned in the 1D protein modality can be adequate for the relatively simple task of affinity prediction. However, it does not generalize well for the relatively difficult task of contact prediction especially when the proteins are new. In other words, the accuracy of learned sequence-contact mapping can be restricted to seen proteins or their homologs but does not transfer to a non-homolog. In contrast, structural topology in the 2D protein modality is more difficult to capture by GNNs and its mapping to affinity can be predicted less well (not to mention that the quality of the predicted 2D modality is worse than the actual). However, once the mapping between the 2D embeddings and intermolecular contacts is

learned, it generalizes well to unseen proteins, possibly due to better capturing subgraph (binding pocket) similarity.

Our experiments also 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. The newly proposed cross interaction model has led to better affinity prediction (RMSE and Pearson's r) and better contact prediction (AUPRC) for unseen proteins than SOTAs, any our single-modality model, and the simple cross-modality model with concatenation. It has also outperformed those other models in the generalizability of binding-site prediction for unseen proteins.

Acknowledgment

This project is in part supported by the National Science Foundation (CCF-1943008 to YS) and the National Institute of General Medical Sciences of the National Institutes of Health (R35GM124952 to YS). We thank Texas A&M High Performance Research Computing (HPRC) for computing allocations. We also thank anonymous reviewers for useful comments that have helped improve the manuscript.

References

[1] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery, 3(8):711–716, 2004.

[2] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery, 9(3):203–214, 2010.

[3] Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. Journal of chemical information and modeling, 59(9):3981–3988, 2019.

[4] W. Torng and R. B. Altman. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. J Chem Inf Model, 59(10):4131–4149, 10 2019.

[5] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. Atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint arXiv:1703.10603, 2017.

[6] J. Jimenez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model, 58(2):287–296, Feb 2018.

[7] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. Bioinformatics, 34(17):i821–i829, 2018.

[8] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In IJCAI, volume 2018, pages 3371–3377, 2018.

[9] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics, 35(18):3329–3338, 2019.

[10] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics, 35(2):309–318, 2019.

[11] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. RSC Advances, 10(35):20701–20712, 2020.

[12] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Explainable deep relational networks for predicting compound-protein affinities and contacts. arXiv preprint arXiv:1912.12553, 2019.

[13] Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: A multi-objective neural network for predicting compound-protein interactions and affinities. Cell Systems, 10(4):308–322, 2020.

[14] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.

[15] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. L^2-gcn: Layer-wise and learned efficient training of graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2127–2135, 2020.

[16] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When does self-supervision help graph convolutional networks? arXiv preprint arXiv:2006.09136, 2020.

[17] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations, 2020.

[18] Meng Liu, Hongyang Gao, and Shuiwang Ji. Towards deeper graph neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 338–348, 2020.

[19] Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, and Jiliang Tang. Adversarial attacks and defenses on graphs: A review and empirical study. arXiv preprint arXiv:2003.00653, 2020.

[20] Salah El Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for long-term dependencies. In Advances in neural information processing systems, pages 493–499, 1996.

[21] Y. Cao and Y. Shen. Energy-based graph convolutional networks for scoring protein docking models. Proteins, 88(8):1091–1099, 08 2020.

[22] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1):e1005324, 2017.

[23] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. Bioinformatics, 31(3):405–412, 2015.

[24] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research, 35(suppl_1):D198–D201, 2007.

[25] Roman A Laskowski, Jagoda Jabłońska, Lukáš Pravda, Radka Svobodová Vařeková, and Janet M Thornton. Pdbsum: Structural summaries of pdb entries. Protein science, 27(1):129–134, 2018.

[26] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Advances in neural information processing systems, pages 1024–1034, 2017.

[27] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826, 2018.

[28] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.