CoPRA: Bridging Cross-domain Pretrained Sequence Models with Complex Structures for Protein-RNA Binding Affinity Prediction

Rong Han[1*], Xiaohong Liu[2*], Tong Pan[3,4], Jing Xu[3,4], Xiaoyu Wang[3,4], Wuyang Lan[5], Zhenyu Li[1], Zixuan Wang[1], Jiangning Song[3,4†], Guangyu Wang[5†], Ting Chen[1†]

[1] BNRist, Department of Computer Science and Technology, Tsinghua University
[2] UCL Cancer Institute, University College London
[3]Monash Data Futures Institute, Monash University
[4] Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University
[5] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
hanr21@mails.tsinghua.edu.cn, xiaohong.liu@ucl.ac.uk,
Jiangning.Song@monash.edu, guangyu.wang24@gmail.com, tingchen@tsinghua.edu.cn

Abstract

Accurately measuring protein-RNA binding affinity is crucial in many biological processes and drug design. Previous computational methods for protein-RNA binding affinity prediction rely on either sequence or structure features, unable to capture the binding mechanisms comprehensively. The recent emerging pre-trained language models trained on massive unsupervised sequences of protein and RNA have shown strong representation ability for various in-domain downstream tasks, including binding site prediction. However, applying different-domain language models collaboratively for complex-level tasks remains unexplored. In this paper, we propose CoPRA to bridge pre-trained language models from different biological domains via Complex structure for Protein-RNA binding Affinity prediction. We demonstrate for the first time that cross-biological modal language models can collaborate to improve binding affinity prediction. We propose a Co-Former to combine the cross-modal sequence and structure information and a bi-scope pre-training strategy for improving Co-Former's interaction understanding. Meanwhile, we build the largest protein-RNA binding affinity dataset PRA310 for performance evaluation. We also test our model on a public dataset for mutation effect prediction. CoPRA reaches state-of-the-art performance on all the datasets. We provide extensive analyses and verify that CoPRA can (1) accurately predict the protein-RNA binding affinity; (2) understand the binding affinity change caused by mutations; and (3) benefit from scaling data and model size.

Code — https://github.com/hanrthu/CoPRA
Extended version — https://arxiv.org/abs/2409.03773

Introduction

Protein-RNA interactions are crucial in various biological processes, including gene expression and regulation (Corley, Burns, and Yeo 2020), protein translocation, and the cell cycle (Zhou et al. 2020). Understanding the mechanism of protein-RNA binding is the cornerstone of unraveling complex gene regulatory processes and deciphering the genetic underpinning of diseases, such as neurodegenerative disorders (Gebauer et al. 2021) and kidney disease (Seufert et al. 2022), leading to the advancement of RNA-based therapies and the design of protein inhibitors that specifically target these interactions. However, protein-RNA binding is highly flexible. Some proteins bind with RNA through canonical regions while others bind with RNA through intrinsically disordered regions - protein domains characterized by low sequence complexity and highly variable structures (Seufert et al. 2022), making it challenging to model the mechanism.

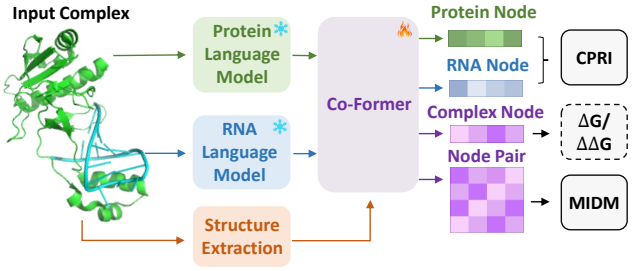

Figure 1: CoPRA combines Protein and RNA language models with structure information by pre-training on bi-scope tasks with different special embeddings. CPRI: Contrastive Protein-RNA interaction modeling; MIDM: Mask interface distance modeling. The dashed line represents downstream tasks.

Several computational methods have been proposed for protein-RNA binding affinity prediction, including sequence-based and structure-based methods. The sequence-based approaches process the protein and RNA sequence separately with different sequence encoders (Yang and Deng 2019a; Pandey et al. 2024), and subsequently model the interactions. However, their performance is often limited because the binding affinity is mainly determined by

*These authors contributed equally.
†Corresponding authors.

the binding interface structure (Deng et al. 2019). Other recent methods are structure-based (Hong et al. 2023; Harini, Sekijima, and Gromiha 2024a), focusing on extracting structural features at the binding interface, such as energy and contact distance. Based on the extracted features, they developed structure-based machine-learning approaches for affinity prediction. However, these methods are highly dependent on feature engineering with limited generalization ability on new samples due to the limited development dataset size.

Recently, many protein language models (PLMs) (Lin et al. 2022; Rao et al. 2021) and RNA language models (RLMs) (Penić et al. 2024; Chen et al. 2022) have been developed, most of which utilize a mask language modeling strategy (Devlin 2018) to pre-train the models with massive unlabeled sequences. They've shown great performance and generalization ability in various downstream tasks. As the 3D structure of protein/RNA is crucial for understanding their functions, combining structure information into the LMs has recently become a new trend. For example, SaProt (Su et al. 2024) pre-trains PLMs with structure information and shows increased performance on different tasks. Instead of adding structure information into pre-training directly, other methods use a much lighter way by combining it with a pre-trained sequence model, such as (Jing et al. 2024), showcasing a strong performance gain compared to the sequence-only counterparts. Most of these models are trained and used in single biological modal tasks (i.e. protein or RNA only).

Although the current works show the prosperous potential of structure-informed biological language models for interaction tasks, there are still few works combining pre-trained models from different biological domains. Integrating pre-trained models for multiperspective information extraction has received much attention recently (Li et al. 2024). Modeling cross-modal complex structures for single-modal LMs requires a suitable model design. In the protein-RNA binding affinity prediction task, one key challenge is the limited size of labeled complex structures, as there are only several datasets that contain a small number of protein-RNA affinity labels, e.g. 135 samples in PRBABv2. Meanwhile, some affinity labels from different datasets may conflict with each other, making it hard to develop and evaluate the models. Therefore, applying different-domain language models collaboratively for complex-level tasks remains less explored.

In this paper, we propose CoPRA, the first attempt to bridge a PLM and an RLM via Complex structure for Protein-RNA binding Affinity prediction, as shown in Figure 1. Specifically, the overall pipeline of CoPRA is: The protein and RNA sequences are first input into a PLM and an RLM, respectively. Then, we select the embeddings from the two LMs' outputs that are at the interaction interface as the sequence embedding for the subsequent cross-modality learning. The structure information is also extracted from the interaction interface as the pair embedding. We design a lightweight Co-Former to bridge the interface sequence embedding from two LMs together with the complex structure information. Co-Former combines the sequence and structure information with a structure-sequence fusion module. We also propose a bi-scope pre-training strategy for Co-Former to model coarse-grained contrastive interaction classification (CPRI) and fine-grained interface distance prediction (MIDM) at **atom-wise precision**[1]. To deal with the lack of a unified labeled standard dataset issue, we curated the largest protein-RNA binding affinity dataset PRA310 from three public datasets and evaluated CoPRA and other models' performance. To further demonstrate CoPRA' ability to understand protein-RNA binding, we adopt it to predict the binding affinity change caused by protein mutation. In summary, our main contributions are listed as follows:

- We propose CoPRA, the first attempt to combine protein and RNA language models with complex structure information for protein-RNA binding affinity prediction.
- We design a Co-Former to bridge the embedding of the interface sequence from two LMs together with the complex structure information and design a bi-scope pre-training method, including CPRI and MIDM for understanding the binding from different aspects. Co-Former is trained on our curated unsupervised dataset PRI30k.
- We curate the largest protein-RNA binding affinity dataset PRA310 from multiple data sources. And evaluate the model's performance on three datasets. CoPRA reaches state-of-the-art performance on multiple datasets, including PRA310 and its subset PRA201 for binding affinity prediction, and a mCSM blind-test set for mutation effect on binding affinity prediction.

Related Work

Protein-RNA Binding Affinity Prediction

Several sequence- or structure-based machine learning-based methods have been applied to predict protein-RNA binding affinity. For example, PNAB (Yang and Deng 2019b) is a stacking heterogeneous ensemble framework based on multiple machine learning methods, e.g. SVR and Random Forest. They manually extract different biochemical features from the protein and RNA sequences. DeePNAP (Pandey et al. 2024) is another sequence-based method, leveraging 1D Convolution networks for feature extraction. PredPRBA and PRdeltaGPred (Deng et al. 2019; Hong et al. 2023) employ interface structure features for better prediction. Besides, PRA-Pred (Harini, Sekijima, and Gromiha 2024b) is a multiple linear regression model, which utilizes protein-RNA interaction information as features in addition to the protein and RNA information. These studies demonstrate that the sequence feature of RNA/protein, and the interface structure feature both contribute to more accurate prediction. However, most of them only employ part of the information, and it is demanding to develop a method to leverage both sequence and interface structure information.

Protein and RNA Language Models

Many efforts have emerged to develop foundation language models to leverage the massive biological sequence data. One of the first papers is ESM-1b (Rives et al. 2021) trained on 250 million protein sequences with a BERT-style strategy. Several other PLMs are proposed and perform well on

[1]The distance of nodes is by the nearest atom between them.

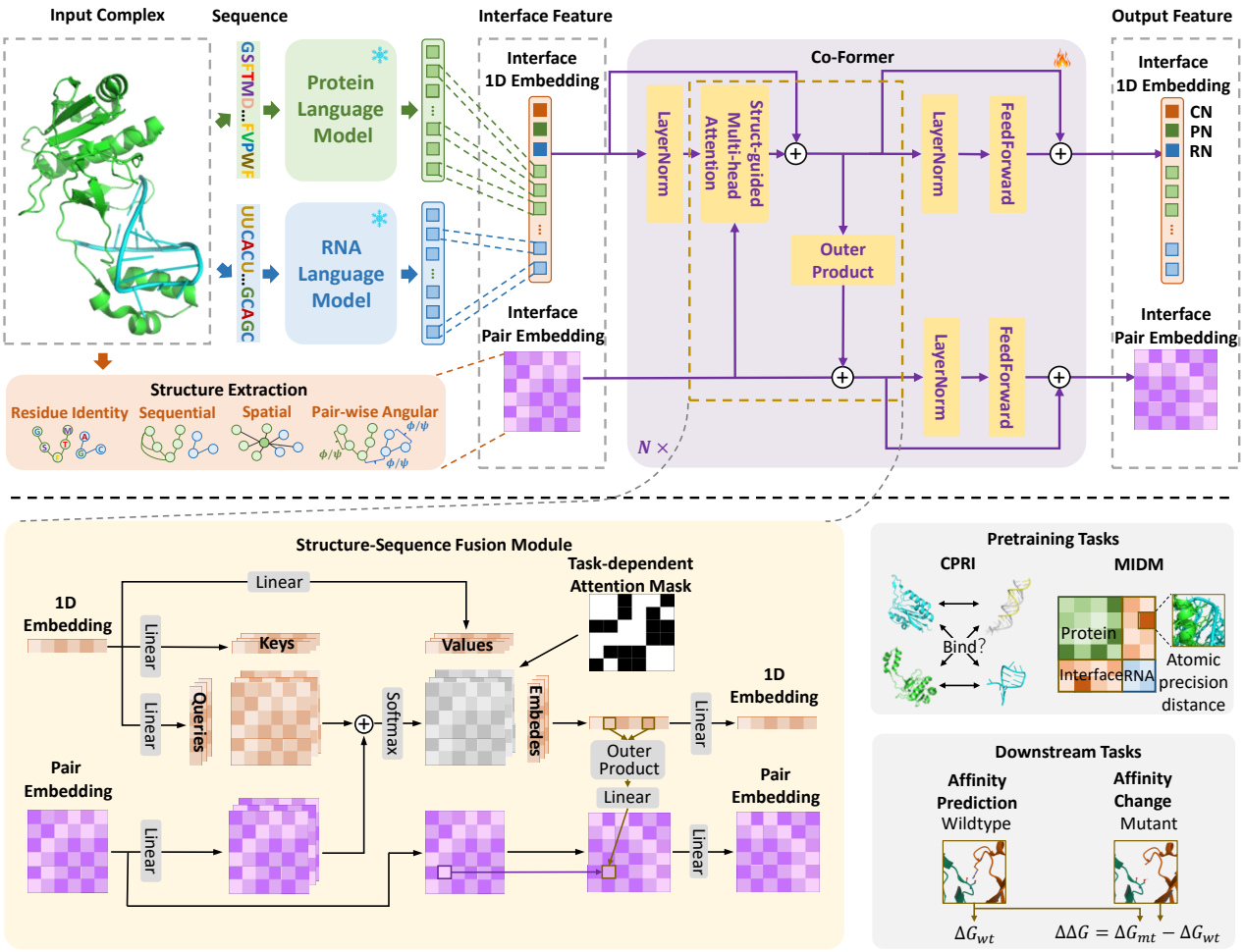

Figure 2: Overview of CoPRA. Given a protein-RNA complex as input, the sequence information are fed into a PLM and an RLM, respectively. The output embeddings at binding interface are fed into Co-Former with pairwise information. The Co-Former fuses the 1D and pair embedding. CN, PN, and RN are special nodes for complex, protein, and RNA, respectively.

various downstream tasks(Rao et al. 2021; Elnaggar et al. 2021; Brandes et al. 2022). Especially, ESMFold (Lin et al. 2022) and OmegaFold (Wu et al. 2022) show the power of PLMs on protein structure prediction, without multiple sequence alignment information as in AlphaFold2 (Jumper et al. 2021). The PLM from ESMFold is named ESM-2, which contains various parameter sizes, from 8M up to 15B. Meanwhile, most RLMs employ a similar paradigm of that in PLMs. The RLMs are trained on massive non-coding RNA sequences. RNA-FM (Chen et al. 2022), Uni-RNA (Wang et al. 2023) and RiNaLMo (Penić et al. 2024) are three representative RLMs. They show great ability in RNA function and secondary structure prediction. While PLMs and RLMs have succeeded in many biological tasks, applying them together remains an unexplored area of research.

Multi-Modal Learning in Language Models

Learning from multiple modals can provide the model with multi-source information of the given context (Huang et al. 2021). Multi-modal learning achieves impressive performance improvement compared to its single-modal counterparts and brings new applications(Luo et al. 2024; Li et al. 2023). Contrastive learning is one efficient unsupervised way to align multi-modal representation to the same semantic space. CLIP (Radford et al. 2021) used an in-batch contrastive strategy to train visual encoders with the text encoders. BLIP-2 (Li et al. 2023) introduces a lightweight QFormer for visual-language pretraining with frozen image encoders and LLMs. In the field of protein, many efforts have been made to integrate the 3D structure information into PLMs. LM-design (Zheng et al. 2023) adds a structure adapter to ESM-2, enabling the structure-informed PLMs on conditional protein design. Recently, SaProt (Su et al. 2024) and ESM-3 (Hayes et al. 2024) pretrain the PLM with protein sequence and its structural information, increasing the models' overall performance. Existing multi-modal PLMs were trained with the protein structure and sequence modals. It is still an open problem for combining multiple biological modals (e.g. protein and RNA) with complex structure information for complex-level interaction tasks.

Methods

In this section, we introduce the details of CoPRA. First, we introduce the overview of CoPRA and some -RNA complex. Next, we present Co-Former for bridging the multi-modal information from protein and RNA. Later, we will describe the pre-training task design, including CPRI and MIDM. At last, we will introduce the formulation of downstream tasks, including binding affinity prediction and mutation effect on affinity change prediction. The overall workflow of CoPRA is described in Figure 2.

CoPRA Overview

CoPRA is designed to leverage the PLM and RLM for binding affinity prediction. Given a protein-RNA complex C, we input the sequence of each protein chain P_i into a PLM, and each RNA chain R_i into an RLM. We generate a sequence embedding and a pair embedding at the binding interface for Co-Former. The Co-Former performs structure-sequence fusion and outputs multi-level representations. To develop Co-Former's multi-modal understanding, we propose a bi-scope pre-training approach, including CPRI and MIDM, enhancing the model's understanding of protein-RNA complex in different granularity.

Notations of the Protein-RNA Complex

The input is a protein-RNA complex with at least one protein chain and one RNA chain each. We define the protein as a set of chains $P = \{P_1, ..., P_n\}$, and RNA as $R = \{R_1, ..., R_n\}$.

Protein. Each protein chain contains 1D sequence information p_i and 3D structure information X_i as input, noted as $P_i = \{p_i, X_i\}$. For a chain of length L_p, we have $p_i \in \mathbb{A}_p^{L_p}$ and $X_i \in \mathbb{R}^{L_p \times k \times 3}$, where $\mathbb{A}_p$ is the alphabet of protein residue types, including 20 normal amino acids and an unknown token 'X'. And k is the number of atoms for representation, we have $k = 4$ for CoPRA modules, containing backbone atoms $\{N, C_A, C, O\}$.

RNA. The input of an RNA chain of length L_r is similar to that of proteins, noted as $R_i = \{r_i, X_i\}$, where $r_i \in \mathbb{A}_r^{L_r}$ and $X_i \in \mathbb{R}^{L_r \times k' \times 3}$. The alphabet $\mathbb{A}_r$ of RNA contains only 4 types of base types $\{A, G, C, U\}$ and an unknown token '_'. Here $k' = 4$ for CoPRA modules, containing backbone atoms $\{P, C_4', C_1', N_1\}$ for pyrimidine base types $\{C, U\}$ and $\{P, C_4', C_1', N_9\}$ for purine base types $\{A, G\}$.

Protein-RNA complex. The protein-RNA complex includes sequence and structure information of each chain, and a complex distance map D, noted as $C = \{P, R, D\}$. $D \in \mathbb{R}^{L \times L}$, where L is the total node number of the complex. D is generated by full-atom geometry to get the precise pair-wise distance between nodes.

Protein-RNA Interface Representation

Given a protein-RNA complex input $C = \{P, R, D\}$, we describe here the process for preparing protein-RNA interface representation for Co-Former. In general, Co-Former takes a mixed representation at the binding interface, noted as $C_I = \{S, Z\}$. $S \in \mathbb{R}^{(n+3) \times d_s}$ and $Z \in \mathbb{R}^{(n+3) \times (n+3) \times d_z}$, where n is the interface size, d_s is the sequence embedding size and d_z is the pair embedding size.

Interface sequence embedding. The full sequence of P and R are fed into PLM and RLM separately to get the full sequence embedding. We select n nodes near the interface according to D. Moreover, we design three special nodes as different-level representation aggregators, including a complex node C^s, a protein node P^s, and an RNA node R^s. C^s can attend to all nodes, while P^s can only attend to nodes from proteins and R^s can only attend to nodes from RNAs. These special nodes are randomly initialized and concatenated in front of the interface node embeddings to form $S = C^s \bigoplus P^s \bigoplus R^s \bigoplus P^n \bigoplus R^n$, where P^n and R^n are embeddings for each protein and RNA node.

Interface structure extraction. Given the interface nodes' (including special nodes) positions of the complex C, we can extract invariant pair-wise structure embeddings for Co-Former. We initialize the special nodes' positions at the geometric center of the interface, the protein, and the RNA, respectively. Inspired by Invariant Point Attention (IPA) in AlphaFold2 (Jumper et al. 2021) for protein feature extraction, we extract four types of pair-wise features from backbone atoms of the complex, including node pair type feature, relative sequential position, distance information, and angular information. The pair-wise information is fed into an embedding layer to form Z. As we take the backbone atom positions, Z is unchanged when mutation affects the sidechain conformation.

Co-Former

Co-Former is an N-block dual-path transformer. Each block contains a Structure-Sequence Fusion module SSF, a layer normalization module LN, and a feed-forward module FFN. The form of the l^{th} block is $\{S^{(l+1)}, Z^{(l+1)}\}$ = $\{\mathrm{FFN}(\mathrm{LN}(\hat{S}^{(l)})), \mathrm{FFN}(\mathrm{LN}(\hat{Z}^{(l)}))\}$, where $\{\hat{S}^{(l)}, \hat{Z}^{(l)}\} = \mathrm{SSF}(\{S^{(l)}, Z^{(l)}\})$. In this section, we will describe the SSF module in detail.

The SSF module consists of two components, a structure-guided multi-head self-attention module and an outer-product update module, as shown in Figure 2. Given the l^{th} layer's input $\{S^{(l)}, Z^{(l)}\}$, the pair embedding $Z^{(l)}$ is first projected to the head size and added to the attention embedding, guiding attention with structural information. Then, we take a pair-wise outer product for the updated sequence embedding $\hat{S}^{(l)}$ to get the pair embedding $\hat{Z}^{(l)}$. The module can be formulated as:

$$Q^{(l)}, K^{(l)}, V^{(l)} = S^{(l)}[W_Q^{(l)}, W_K^{(l)}, W_V^{(l)}], \tag{1}$$

$$A^{(l)} = \frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}} + \mathrm{Linear}(Z^{(l)}), \tag{2}$$

$$\hat{S}^{(l)} = (\mathrm{Softmax}(A^{(l)}) \odot M) \cdot V^{(l)}, \tag{3}$$

$$o_{ij}^{(l)} = \hat{s}_i^{(l)} \otimes \hat{s}_j^{(l)T}, \tag{4}$$

$$\hat{z}_{ij}^{(l)} = z_{ij}^{(l)} + \mathrm{Linear}(\mathrm{o}_{\mathrm{ij}}^{(\mathrm{l})}), \tag{5}$$

where, W_Q, W_K and W_V are the projection matrices, and M is the task-dependent mask. $\hat{s}_i^{(l)}, \hat{s}_j^{(l)} \in \mathbb{R}^{1\times d_s}$ are the i^{th} and j^{th} feature of $\hat{S}^{(l)}$. $\hat{s}_i^{(l)} \otimes \hat{s}_j^{(l)T}$ is the outer product, resulting in $o_j^{(l)} \in \mathbb{R}^{d_s\times d_s}$. $z_{ij}^{(l)}$, $\hat{z}_{ij}^{(l)}$ is from position (i, j) of $Z^{(l)}$ and $\hat{Z}^{(l)}$, respectively. We simplify the multi-head attention in the equation for easy understanding.

Bi-scope Pre-training

In this section, we will describe the pre-training tasks, including a cross-modal contrastive protein-RNA interaction (CPRI) task for understanding interaction pairs (whether they interact) and a mask interface distance modeling (MIDM) for understanding the atom-precision node distance (how they interact) given only backbone structure as input.

Contrastive interaction modeling. Utilizing protein and RNA representations for cross-modal matching is similar to image-text matching. We formulate this problem in an in-batch way, inspired by CLIP (Radford et al. 2021). Specifically, for a protein P and an RNA R from a complex C, we mask the interface structure information of the pair embedding Z and get the output protein and RNA special node embedding from Co-Former, denoted as P^s, R^s. Given a batch of protein-RNA complexes of batch size K, we generate K^2 pairs (P_i^s, R_j^s), where $i, j \in \{1, ..., K\}$. The pair is positive when $i = j$ and the other pairs are negative. We adopted a symmetric contrastive loss function for the training:

$$\mathcal{L}_i^P(P_i^s, \{R_j^s\}_{j=1}^K) = -\frac{1}{K}\log\frac{\exp(s(P_i^s, R_i^s)/\tau)}{\sum_j \exp(s(P_i^s, R_j^s)/\tau)}, \quad (6)$$

$$\mathcal{L}_i^R(R_i^s, \{P_j^s\}_{j=1}^K) = -\frac{1}{K}\log\frac{\exp(s(R_i^s, P_i^s)/\tau)}{\sum_j \exp(s(R_i^s, P_j^s)/\tau)}, \quad (7)$$

$$\mathcal{L}_{\mathcal{CPRI}} = \frac{1}{2}\sum_{i=1}^{K}(\mathcal{L}_i^P + \mathcal{L}_i^R), \quad (8)$$

where, s denotes the similarity of the embeddings and we adopt cosine similarity in practice, and τ is the temperature.

Mask interface modeling. Modeling the atom-precision distance is crucial for understanding how the protein-RNA nodes interact. We design a coarse- to fine-grained pre-training method. Specifically, 50% of the pair embedding Z will be masked with a ratio of 15%, and the other 50% will be unchanged. The model is required to reconstruct the interface distance. The ground truth distance of two nodes is defined by the nearest atoms from each node, thus the model needs to infer the interface detail from sequence embedding and partially masked pair embedding Z. **All** the distance at the **interface** will be used for loss calculation. To make the training more stable, we divide the distance into multiple bins, where the bins at the close part are dense and at the remote part are sparse, converting the task into a classification task with a cross-entropy loss:

$$O_i = \text{Interface}(\text{Linear}(Z_i^{(N)})), \quad (9)$$

$$\mathcal{L}_{MIDM,i} = -\frac{1}{L^2}\sum_{j,k=1}^{L}\log\frac{\exp(o_{ijk,t}/\tau)}{\sum_b \exp(o_{ijk,b}/\tau)}y_{ijk,t}, \quad (10)$$

where, O_i is the distance prediction of the i^{th} complex, and $y_{ijk,t}$ is the distance for the i^{th} complex at position (j, k), with the label t, and τ is the temperature. With a hyperparameter α, the in-batch bi-scope pre-training loss is :

$$\mathcal{L} = \mathcal{L}_{CPRI} + \alpha \cdot (\frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_{MIDM,i}). \quad (11)$$

Protein-RNA Affinity Prediction Tasks

The downstream tasks consist of protein-RNA binding affinity prediction and protein mutation effect on binding affinity prediction. Here is the formulation of these two tasks. We take MSE loss for both tasks.

Binding affinity prediction. Given a complex C as input, we fed the output special node's embedding C^s into an MLP to predict ΔG, noted as $\Delta G = \text{MLP}(C^s)$.

Mutation effect on binding affinity prediction . This task predicts the binding affinity change between the mutant and the wild complex[2], noted as $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wild}$. Since Co-Former only requires backbone structure information, we can input the same backbone structure and different sequences to get C_{wild}^s, C_{mut}^s, making it convenient for prediction, note as $\Delta\Delta G = \text{MLP}(C_{mut}^s) - \text{MLP}(C_{wild}^s)$.

Experiments

Exeriment Setup

Pre-training dataset. The pre-training dataset used here is curated by ourselves, capturing protein-RNA pairs of multiple poses. There are in total 5,909 protein-RNA complexes in the Protein Data Bank (PDB), which were collected in a pair-wise form in BioLiP2. They define each interacting protein-RNA chain pair in the complex as an entry, resulting in 150k chain pairs. We create the non-redundant pre-training dataset PRI30k with the annotation of BioLiP2 by finding the maximum connected subgraph in each complex.

Affinity datasets. Existing affinity datasets only contain a small number of protein-RNA affinity data with inconsistent labels across datasets. It is necessary to build a standard dataset for benchmarking. We collect samples from three public datasets, PDBbind (Wang et al. 2004), PRBABv2 (Hong et al. 2023), and ProNAB (Harini et al. 2022). After removing duplication we get 435 unique complexes. We carefully compare the inconsistent labels from the raw literature and calibrate the annotations. We then filter complexes with length and chain number limits, resulting in 310 complexes. We name our dataset PRA310, which is the largest and most reliable dataset under the same settings. We utilize CD-HIT (Fu et al. 2012) to generate complex clusters, with a sequence identity threshold of 70%. We split these clusters for a standard 5-fold cross-validation setting. PRA201 is a subset of PRA310, containing only one protein chain and one RNA chain in each complex with a stricter length limit. The mCSM blind test set is a dataset from mCSM (Pires,

[2]This is the common representation, while in mCSM, the label is defined as $\Delta\Delta G = \Delta G_{wild} - \Delta G_{mut}$.

Method	Struc	Seq	LM	PRA310				PRA201			
				RMSE↓	MAE↓	PCC↑	SCC↑	RMSE↓	MAE↓	PCC↑	SCC↑
LM+LR	✗	✓	✓	1.801	1.472	0.365	0.348	1.750	1.383	0.370	0.362
LM+RF	✗	✓	✓	1.561	1.248	0.418	0.457	1.569	1.252	0.437	0.467
LM+MLP	✗	✓	✓	1.688	1.388	0.412	0.428	1.638	1.282	0.403	0.412
LM+SVR	✗	✓	✓	1.506	1.209	0.475	0.489	1.476	1.192	0.454	0.456
LM+Transformer	✗	✓	✓	1.481	1.192	0.489	0.485	1.433	1.172	0.492	0.487
DeepNAP* (Pandey et al. 2024)	✗	✓	✗	-	-	-	-	1.964	1.600	0.345	0.349
PredPRBA* (Deng et al. 2019)	✓	✗	✗	-	-	-	-	2.238	1.695	0.370	0.316
FoldX† (Delgado et al. 2019)	✓	✗	✗	-	-	0.212	0.283	-	-	0.212	0.268
GCN (Kipf and Welling 2016)	✓	✗	✗	1.705	1.378	0.145	0.144	1.631	1.322	0.201	0.203
GAT (Veličković et al. 2017)	✓	✗	✗	1.644	1.337	0.238	0.174	1.542	1.235	0.262	0.221
EGNN (Satorras et al. 2021)	✓	✗	✗	1.634	1.340	0.226	0.212	1.639	1.345	0.241	0.217
GVP (Jing et al. 2020)	✓	✗	✗	1.678	1.361	0.262	0.283	1.702	1.372	0.240	0.305
IPA (Jumper et al. 2021)	✓	✗	✗	1.462	1.208	0.495	0.496	1.464	1.191	0.510	0.514
LM+IPA	✓	✗	✓	1.454	1.198	0.514	0.496	1.405	1.159	0.532	0.507
CoPRA (scratch)	✓	✓	✓	1.446	1.188	0.522	0.520	1.428	1.172	0.534	0.526
CoPRA	✓	✓	✓	**1.391**	**1.129**	**0.580**	**0.589**	**1.339**	**1.059**	**0.569**	**0.587**

Table 1: The mean metrics of 5-fold cross-validation on the PRA310 and PRA201 datasets. * They only provide web servers with restrictions, so we only test them on PRA201. † The FoldX prediction is the energy change, thus we only compare the correlation coefficient. LM is ESM-2 + RiNALMo. CoPRA (scratch) represents CoPRA not pre-training with CPRI and MIDM.

Ascher, and Blundell 2014), containing 79 non-redundant single-point mutations from 14 protein-RNA complexes.

Metrics and implementation details. Following (Pandey et al. 2024), we take 4 metrics for evaluation, including the root mean square error (RMSE), the mean absolute error (MAE), the Pearson correlation coefficient (PCC), and the Spearman correlation coefficient (SCC). We take ESM-2 650M (Lin et al. 2023) and RiNALMo 650M (Penić et al. 2024) as our LMs. All the experiments are conducted on 4 NVIDIA A100-80G GPUs. The block number of Co-Former is 6, with a sequence and pair embedding size of 320 and 40, respectively. In pre-training, we set MIDM's mask ratio to 15%. We use the Adam optimizer with an initial learning rate of 3e-5. The node number of the interface is 256. The introduction of baselines is in the extended version.

Predicting Protein-RNA Binding Affinity

We first evaluate our model's performance on PRA310 and PRA201. We divide the baseline methods into sequence- and structure-based. As illustrated in Table 1, the scratch version of CoPRA reaches the best performance on the PRA310 dataset. IPA is the best-performed model without LMs, and we replace the sequential input of IPA with the embeddings from LMs, improving its performance with 0.19 on PCC. Moreover, most methods with LM embedding as input perform better than others, indicating the great power of combining pre-trained unimodal LMs for affinity prediction. We then pre-train our model with PRI30k, increasing the overall performance significantly on both datasets. On PRA310, CoPRA gets an RMSE of 1.391, MAE of 1.129, PCC of 0.580, and SCC of 0.589, much better than the second-best model CoPRA (scratch). The PredPRBA and DeepNAP only provide web servers and support protein-RNA pair affinity prediction, and we compared the methods on the PRA201 dataset with them. Although at least 100 samples in PRA201 appear in their training set, their performance on PRA 201 is significantly lower than that they reported, indicating the less generalization ability of these methods. This phenomenon can be explained by the experiment of PRdeltaGPred (Hong et al. 2023) that removes worst-performed samples. Moreover, we observe a consistent performance improvement of most models from PRA310 to PRA201, indicating that PRA310 is more challenging. The experiments in PRA310 and PRA201 show CoPRA's ability to precisely predict the binding affinity, especially when equipped with the proposed bi-scope pre-training.

Method	RMSE↓	MAE↓	PCC↑	SCC↑
FoldX (zero-shot)	1.727	1.496	**0.474**	**0.548**
CoPRA (zero-shot)	**0.994**	**0.737**	0.314	0.411
DeepNAP*	1.106	1.004	0.428	0.339
mCSM	1.814	1.478	0.528	0.466
CoPRA	**0.957**	**0.833**	**0.550**	**0.570**

Table 2: Per-structure performance on mCSM blind test set. * DeePNAP's training set overlaps with this test set.

Predicting Mutation Effects on Binding Affinity

To further evaluate our model's understanding of affinity in a fine-grained way, we redirect our model to predict the protein's single-point mutation effect on the protein-RNA complex. Following works in protein mutation effects prediction (Luo et al. 2023), the metrics are averaged at a **per-complex** level. We evaluate both zero-shot and fine-tuned performance of CoPRA, after pre-training on PRI30k and tuning on PRA310. As shown in Table 2, Notably, ours (zero-shot)

has a competitive performance, outperforming other models under the RMSE and MAE metrics. After fine-tuning with the cross-validation set used by mCSM, our model outperforms other models in all four metrics, with RMSE of 0.957, MAE of 0.833, PCC of 0.550, and SCC of 0.570. This superior performance comes from the bi-scope pre-training targets, although not see any mutational complex structures. The performance demonstrates CoPRA's generalization ability on different affinity-related tasks.

Ablation Study

In this section, we present extensive ablation studies of our model to explore its performance on PRA310, including the module parts, the pertaining strategy, and the model size.

Method	RMSE↓	MAE↓	PCC↑	SCC↑
CoPRA	**1.391**	**1.129**	**0.580**	**0.589**
- Pre-train	1.446	1.188	0.522	0.520
- Pair info	1.454	1.177	0.518	0.519
- Crop patch	1.481	1.192	0.489	0.485
- Special nodes	1.497	1.211	0.456	0.469
- Co-Former	1.688	1.388	0.412	0.479

Table 3: Ablation study on modules

Modules ablation. We **progressively delete** the modules of CoPRA. As shown in Table 3, removing each component of CoPRA will cause a performance decrease, demonstrating the necessity and importance of the modules we designed. The removal of pre-training causes a significant loss of performance, indicating that our pre-training strategy is crucial for affinity prediction. However, the removal of pair information from the scratch version of CoPRA does not cause a significant loss of performance while removing the patch cropping will cause an obvious decrease. Because the interface information can help the model directly, adding more information on top of the interface cropping may not be helpful when the sample number is limited and the binding mode is flexible. The special nodes also increase the model's performance because they are indeed different levels of attention-based readout functions, effective for multi-level representation of the complex. If we remove all the components and only feed the LMs' output into an MLP, the performance will be much poorer, thus brutally combining embeddings without a suitable model is impracticable.

Method	RMSE↓	MAE↓	PCC↑	SCC↑
Scratch	1.446	1.188	0.522	0.520
CPRI	1.442	1.165	0.528	0.522
DM	1.445	1.167	0.542	0.535
CPRI+DM	1.418	1.167	0.558	0.541
CPRI+IDM	1.421	1.142	0.560	0.542
CPRI+MIDM	**1.391**	**1.129**	**0.580**	**0.589**

Table 4: Ablation study on pretraining strategy

Pretraining strategy ablation. Based on Table 4, we can observe that when only trained with one pre-training target, the distance modeling (DM) brings better performance than CPRI. The distance modeling task is more fine-grained and provides more information for affinity tasks. Combining CPRI with various DM tasks improves the overall performance. Moreover, the results suggest that distance at the interface is more important than that within protein and RNA, thus directly modeling the interface is better. After masking some of the pair embeddings, the task becomes more challenging, urging the model to get an in-depth understanding of the relationship between the node type and distance.

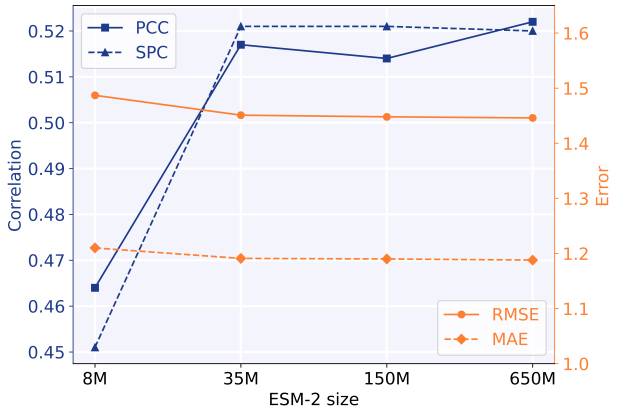

Figure 3: Ablation study of ESM-2 model size.

Model size ablation Since RiNALMo only provides a 650M pre-trained model, we ablate the size of ESM-2 and train CoPRA from scratch. As shown in Figure 3, increasing the model size brings improvement in performance, and the best-performed model is the ESM-2 650M model. Larger pre-trained models can provide larger embedding dims, containing better representation ability gained from unsupervised sequences. This is consistent with the performance trends observed in the unimodal models. We demonstrate that when doing cross-modal tasks, the collaborator model's size is also of important consideration, and the larger model will probably result in better complex-level performance.

Conclusion

In this work, we present CoPRA, the first attempt to combine different biological language models with structural information for protein-RNA binding affinity prediction. We design a Co-Former for sequence and structure feature fusion and propose an effective bi-scope pre-training approach. Meanwhile, we curate the largest standard protein-RNA binding affinity dataset PRA310, and a pre-training dataset PRI30k. Our model achieves state-of-the-art performance on binding affinity and mutation effect prediction.

Future research can be applied to more biological domains, such as protein-DNA binding. While our model performs well in predicting the protein's single-point mutation effect on the complex, it is also important to extend the application to multi-point mutation and RNA mutations.

Acknowledgments

This work is supported by the National Key R&D Program of China (2021YFF1201300, 2021YFF1201303, 2022YFC2703105), National Natural Science Foundation of China (grants 62272055), New Cornerstone Science Foundation through the XPLORER PRIZE, Guoqiang Institute of Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist). We also acknowledge financial support from the Major and Seed Inter-Disciplinary Research projects awarded by Monash University (J.S.). The funders had no roles in study design, data collection and analysis, the decision to publish, or manuscript preparation.

References

Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; and Linial, M. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8): 2102–2110.

Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T.; et al. 2022. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*.

Corley, M.; Burns, M. C.; and Yeo, G. W. 2020. How RNA-binding proteins interact with RNA: molecules and mechanisms. *Molecular cell*, 78(1): 9–29.

Delgado, J.; Radusky, L. G.; Cianferoni, D.; and Serrano, L. 2019. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20): 4168–4169.

Deng, L.; Yang, W.; Liu, H.; and Zhang, S.-W. 2019. Pred-PRBA: Prediction of Protein-RNA Binding Affinity Using Gradient Boosted Regression Trees. *Frontiers in Genetics*, 10: 637.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23): 3150–3152.

Gebauer, F.; Schwarzl, T.; Valcárcel, J.; and Hentze, M. W. 2021. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics*, 22(3): 185–198.

Harini, K.; Sekijima, M.; and Gromiha, M. M. 2024a. PRA-Pred: Structure-based prediction of protein-RNA binding affinity. *International Journal of Biological Macromolecules*, 259: 129490.

Harini, K.; Sekijima, M.; and Gromiha, M. M. 2024b. PRA-Pred: Structure-based prediction of protein-RNA binding affinity. *International Journal of Biological Macromolecules*, 259: 129490.

Harini, K.; Srivastava, A.; Kulandaisamy, A.; and Gromiha, M. M. 2022. ProNAB: database for binding affinities of protein–nucleic acid complexes and their mutants. *Nucleic Acids Research*, 50(D1): D1528–D1534.

Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; and Rives, A. 2024. Simulating 500 million years of evolution with a language model.

Hong, X.; Tong, X.; Xie, J.; Liu, P.; Liu, X.; Song, Q.; Liu, S.; and Liu, S. 2023. An updated dataset and a structure-based prediction model for protein–RNA binding affinity. *Proteins: Structure, Function, and Bioinformatics*, 91(9): 1245–1253. Publisher: John Wiley & Sons, Ltd.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Jing, B.; Eismann, S.; Suriana, P.; Townshend, R. J. L.; and Dror, R. 2020. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*.

Jing, L.; Xu, S.; Wang, Y.; Zhou, Y.; Shen, T.; Ji, Z.; Fang, H.; Li, Z.; and Sun, S. 2024. CrossBind: Collaborative Cross-Modal Identification of Protein Nucleic-Acid-Binding Residues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 2661–2669. Number: 3.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Z.; Zhang, Y.; Pan, T.; Sun, Y.; Duan, Z.; Fang, J.; Han, R.; Wang, Z.; and Wang, J. 2024. FocusLLM: Scaling LLM's Context by Parallel Decoding. *arXiv preprint arXiv:2408.11745*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022: 500902.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al.

2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Luo, L.; Tang, B.; Chen, X.; Han, R.; and Chen, T. 2024. VividMed: Vision Language Model with Versatile Visual Grounding for Medicine. *arXiv preprint arXiv:2410.12694*.

Luo, S.; Su, Y.; Wu, Z.; Su, C.; Peng, J.; and Ma, J. 2023. Rotamer Density Estimator is an Unsupervised Learner of the Effect of Mutations on Protein-Protein Interaction.

Pandey, U.; Behara, S. M.; Sharma, S.; Patil, R. S.; Nambiar, S.; Koner, D.; and Bhukya, H. 2024. DeePNAP: A Deep Learning Method to Predict Protein–Nucleic Acid Binding Affinity from Their Sequences. *Journal of Chemical Information and Modeling*, 64(6): 1806–1815. Publisher: American Chemical Society.

Penić, R. J.; Vlašić, T.; Huber, R. G.; Wan, Y.; and Šikić, M. 2024. RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks. ArXiv:2403.00043 [cs, q-bio].

Pires, D. E.; Ascher, D. B.; and Blundell, T. L. 2014. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3): 335–342.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; and Rives, A. 2021. MSA transformer. In *International Conference on Machine Learning*, 8844–8856. PMLR.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Satorras, V. G.; Hoogeboom, L. G.; Cianferoni, D.; and Serrano, L. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*, 9323–9332. PMLR.

Seufert, L.; Benzing, T.; Ignarski, M.; and Müller, R.-U. 2022. RNA-binding proteins and their role in kidney disease. *Nature Reviews Nephrology*, 18(3): 153–170.

Su, J.; Li, Z.; Han, C.; Zhou, Y.; Shan, J.; Zhou, X.; Ma, D.; The OPMC; Ovchinnikov, S.; and Yuan, F. 2024. SaprotHub: Making Protein Modeling Accessible to All Biologists.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, R.; Fang, X.; Lu, Y.; and Wang, S. 2004. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12): 2977–2980.

Wang, X.; Gu, R.; Chen, Z.; Li, Y.; Ji, X.; Ke, G.; and Wen, H. 2023. UNI-RNA: universal pre-trained models revolutionize RNA research. *bioRxiv*, 2023–07.

Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; et al. 2022. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, 2022–07.

Yang, W.; and Deng, L. 2019a. PNAB: prediction of protein-nucleic acid binding affinity using heterogeneous ensemble models. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 58–63. IEEE.

Yang, W.; and Deng, L. 2019b. PNAB: Prediction of protein-nucleic acid binding affinity using heterogeneous ensemble models. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 58–63.

Zheng, Z.; Deng, Y.; Xue, D.; Zhou, Y.; Ye, F.; and Gu, Q. 2023. Structure-informed language models are protein designers. In *International conference on machine learning*, 42317–42338. PMLR.

Zhou, W.-Y.; Cai, Z.-R.; Liu, J.; Wang, D.-S.; Ju, H.-Q.; and Xu, R.-H. 2020. Circular RNA: metabolism, functions and interactions with proteins. *Molecular cancer*, 19: 1–19.