Enhancing RNA-ligand affinity prediction using transfer learning from large-scale protein-ligand synthetic data

Group Members

Jinhang Wei, Yangcheng Zhang

Project contributions:

Jinhang Wei: conceptual design of the transfer learning framework, reproduction of the model, and the writing of the Abstract and Results sections.

Yangcheng Zhang: performed the data processing, conducted the baseline comparisons against AutoDock Vina and RSAPred, and contributed to the background research and figure generation. Both authors collaborated on the final analysis.

Abstract

Predicting the binding affinity between RNA and small molecule ligands represents a crucial challenge in modern drug discovery, essential for developing novel therapeutics like Risdiplam. However, the development of accurate structure-based predictive models is significantly impeded by the scarcity of experimental RNA-ligand structural data, with fewer than 150 experimentally resolved complexes currently accessible in public databases (PDBbind-RNA[1]). While geometric deep learning has revolutionized protein-ligand interaction modeling, its application to RNA is constrained by this data gap. Recently, some studies have analyzed the BindingDB[2] and ChEMBL[3] databases and found that many experimentally measured affinity values lack corresponding 3D structural information. Therefore, they generated millions of synthetic protein-ligand complexes using the Boltz model and conclusively proven experimentally that these synthetic structures can effectively enhance protein-ligand affinity prediction. In this work, we explore these models parameters trained on large-scale protein-ligand synthetic data for transfer learning, fine-tuning them on an RNA-ligand affinity dataset to explore the feasibility of this approach. The results show that our model, GatorAffinity-RNA, achieves state-of-the-art performance on the PDBbind-RNA dataset, with an average RMSE of 1.290 and a Pearson correlation coefficient of 0.611 in five-fold cross-validation, outperforming traditional molecular docking methods like AutoDock Vina[4] and other deep learning models[5], [6].

Introductionn

In drug development, the affinity between a drug and its target determines the strength and efficiency with which a candidate drug molecule binds to the target, thus influencing lead compound screening, optimization, and druggability. Previously, the vast majority of drugs were designed to bind to protein targets, but in recent years, an increasing number of small molecule drugs have been designed to target RNA, from classic antibiotics to the first approved RNA small molecule drug, risdiplam[7]. This has made RNA-ligand (drug) affinity prediction a crucial factor in drug discovery: applying a highly accurate affinity prediction model during the screening phase allows for the prioritization of compounds with stronger binding affinity and more pronounced mechanisms of action, significantly reducing the time and economic costs of biochemical screening, drastically minimizing the use of ineffective compounds, and lowering the risk of the drug development pipeline. While RNA-ligand affinity prediction models hold significant promise, the highly flexible and multiconformal nature of RNA structures (the same RNA molecule exhibiting vastly different folding states under varying conditions) makes it difficult to obtain stable and reliable 3D structures of RNA-ligand complexes experimentally. This results in a severe scarcity of structural data for RNA-ligand tasks, hindering the acquisition of sufficient training data to improve the generalization ability of affinity prediction models. Existing work attempts to mitigate this data scarcity using pure sequence models and pre-training methods, but these efforts, while effective, are limited. For example, sequence methods neglect crucial 3D geometric constraints, making it difficult to systematically express fine-grained interactions such as hydrogen bond networks, stacking, and hydrophobic pockets. The lack of such physical constraints severely limits the model's performance ceiling. Furthermore, although pre-training is widely used to alleviate the small data problem, in RNA-ligand affinity prediction, relying solely on pre-training often fails to provide decisive improvements. The fundamental reason is that most pre-training tasks learn "general representations," while affinity prediction relies on highly structured and physically constrained 3D interaction patterns. When there is a lack of sufficient paired data of "RNA-ligand complex structure + experimental affinity" downstream, even models with stronger representational capabilities struggle to be adequately calibrated to these realistic conformational affinity relationships, resulting in limited generalization ability. Therefore, RNA-ligand affinity prediction requires a training strategy that can effectively learn 3D interaction patterns under data-scarce conditions and stably transfer to RNA scenarios. Based on this motivation, we propose a strategy to transfer knowledge from large-scale protein-ligand complex-affinity data folded by Boltz models to RNA-ligand affinity tasks. Since both tasks use the same affinity metric (e.g., Kd) and can be unified to the atomic level in input representation, the model has the opportunity to learn affinity prediction patterns universally applicable to biomolecules. Therefore, we used an SE(3)-equivariant structure-based affinity model (GatorAffinity[8]) as the baseline model and pre-trained its parameters on large-scale protein-ligand synthesis complex data. We then transferred this model to the RNA-ligand scenario and fine-tuned it on limited experimental RNA-ligand labeled data to adapt it to the RNA system. For evaluation, we used a cross-validation and ablation experiment system to test the effectiveness of transfer learning: by comparing the model with and without protein-ligand pre-training, and the differences with docking and sequence baseline methods, we verified whether the affinity knowledge learned in the pre-training stage could bring stable gains in the RNA-ligand task.

Contributions:

- Outperforms existing baseline methods in both error and correlation metrics.
- Demonstrates the effectiveness of protein-ligand structure-affinity knowledge transfer even under conditions of scarce RNA data.
- More accurate affinity prediction can improve early screening efficiency and reduce experimental and R&D costs.

Background

In drug development, the binding affinity between small molecules and their targets directly determines whether candidate molecules can exert their effects with sufficient strength and efficiency, thus affecting lead compound screening, subsequent optimization, and drug development. In recent years, the targets of small molecule drugs have gradually expanded from traditional proteins to RNA: from classic antibiotics to the first approved RNA-targeted small molecule drug, risdiplam, all demonstrate the clear drug discovery value of RNA targets. However, in practice, affinity verification mainly relies on low-throughput biochemical experiments, which are costly and time-consuming. Therefore, if reliable RNA-ligand affinity prediction models can be used for sorting and filtering during the screening stage, stronger and more promising compounds can be prioritized, thereby reducing the time and economic cost of experimental screening. An additional challenge in RNA-ligand tasks lies in the "structural data bottleneck." RNA often exhibits greater conformational flexibility and multiple conformations, making it more difficult to obtain stable and reproducible 3D structures of RNA-ligand complexes experimentally, while structural models require paired data of "complex structure + experimental affinity tags" for training. Correspondingly, there is currently very little experimental RNA-ligand data available for fine-tuning; for example, there are only 143 samples for PDBBind-RNA[1]. This has resulted in RNA-specific structural models remaining in a data-constrained state for a long time, making it difficult to improve their generalization ability.

For RNA-ligand affinity prediction, existing studies typically employ the following methods, which correspond to the main baselines in our poster.
Docking Methods[4]: Docking methods search for binding conformations and estimate affinity using a scoring function. Their advantages include a low deployment threshold and no need for large amounts of labeled data. However, their performance is poor in RNA-ligand systems. The core reason is that their scoring function is a highly approximate empirical model: the official documentation explicitly states that the scoring function is inexact and ignores the user-provided partial charge in energy assessment. This is unfavorable for highly negatively charged nucleic acid systems, thus introducing unavoidable errors.

Self-supervised/Contrastive Learning Pre-Training Methods[6]: These models learn representations of RNA pockets and ligands in three dimensions, followed by affinity regression on limited labeled data. While this "pre-training -> fine-tuning" paradigm can alleviate the small sample size problem to some extent, it is still limited in RNA scenarios by the scarcity and narrow distribution of structural data—pre-training signals often fail to cover the conformational flexibility of RNA, resulting in limited predictive generalization ability of the model across target sites or out-of-distribution samples.

Sequence-Based Methods[5]: Sequence methods can operate even with missing structures, but due to the lack of atomic-level 3D geometric constraints, they struggle to fully express key interactions such as hydrogen bond networks and base stacking. Therefore, their affinity prediction is often limited by their ability to express physical mechanisms.

Dataset

In this work, we leverage three main datasets: two large-scale protein–ligand datasets (SAIR and GatorAffinity-DB) for pre-training, and a specialized RNA–ligand dataset from PDBbind for fine-tuning and evaluation. These datasets provide the structural and biochemical data needed to train and validate our model. Tabel 1 summarizes the key characteristics of each dataset, and detailed descriptions follow.

Dataset	Protein-Ligand affinity pairs	RNA-Ligand affinity pairs	usage
SAIR[9]	1,048,857	N/A	Pretraining
GatorAffinity-DB[8]	456,526	N/A	Pretraining
PDBbind-RNA[1]	N/A	143	Fine-tuning

Tabel 1 Dataset Summary

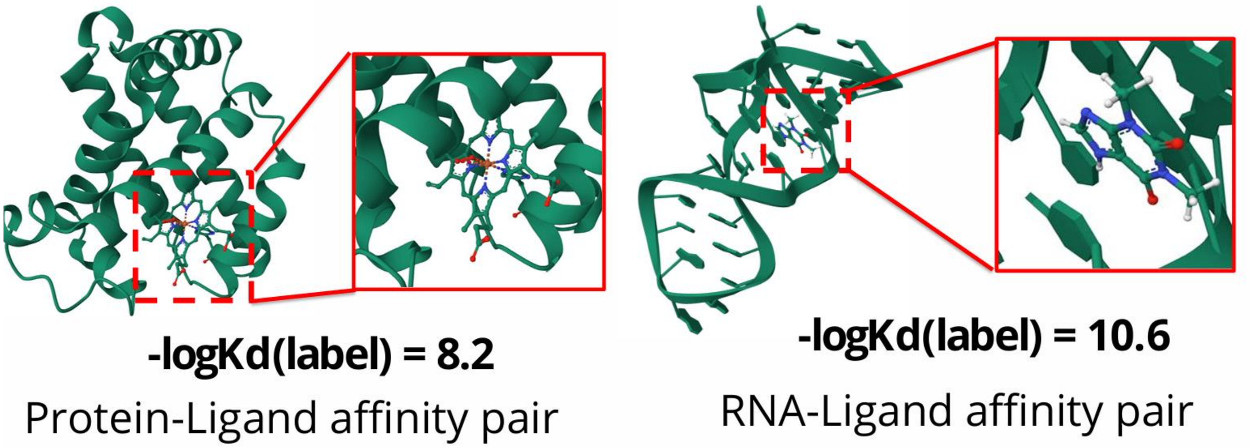


Figure 1 Data format: Protein-ligand complex structure-affinity pairs and RNA-ligand complex structure-affinity pairs

SAIR Dataset

The SAIR dataset (Structurally Augmented IC50 Repository) is the largest publicly available collection of protein–ligand structures with paired potency measurements. It was curated by SandboxAQ in 2025 by mining binding activity data from ChEMBL and BindingDB, and then predicting the corresponding complex structures using a co-folding approach. Specifically, SAIR contains over 1,048,000 unique protein–ligand pairs (with a total of ~5.2 million 3D conformations) derived from experimental affinity data. Each protein–small molecule pair in SAIR has an estimated 3D bound structure, generated using the Boltz-1x model, since most did not have an existing crystal structure. The dataset thus bridges the gap between ligand activities and structures by providing synthetic complex structures labeled with experimental potency values.

GatorAffinity-DB

GatorAffinity-DB is another large-scale protein ligand dataset that focuses more on complexes with quantitative affinity labels such as Kd and Ki. The original binding constants were obtained from publicly available binding databases, and subsequently, several candidate complex structures were generated for each protein ligand pair using structural prediction tools, and the higher quality one was selected as the final structure. The dataset has a scale of several hundred thousand and also contains paired information of "protein structure+ligand structure+affinity label".

Compared to SAIR, GatorAffinity-DB provides a large number of tags with Kd/Ki directly reflecting binding affinity, which are closer to the supervisory signals we ultimately use on RNA. Therefore, we will combine SAIR and GatorAffinity-DB for pre training: SAIR provides a wider structure activity distribution, while GatorAffinity-DB provides quantitative supervision closer to affinity, both of which together help the model learn "atomic level affinity knowledge" in the protein ligand space.

PDBbind-RNA

PDBbind is a database extracted from PDB containing "biomacromolecule-small molecule complexes + literature affinity tags," with only a small subset consisting of RNA-small molecule complexes. This subset is precisely the task we are interested in: the affinity of small molecules binding to various RNA structures.

Compared to the protein portion, the number of RNA-ligand samples is very small, with only about a hundred complexes available. Therefore, it mainly serves two roles in this project:

- As fine-tuning data for downstream tasks: After completing protein-ligand pre-training, we fine-tuned the model on the PDBbind RNA subset to adapt it to the unique backbone structure and other characteristics of RNA.
- As final evaluation data: We performed cross-validation on this dataset to evaluate the model's performance on real-world RNA-ligand affinity prediction tasks.

Methodology

The core idea of this work is to directly adopt the publicly available GatorAffinity structure as the basic model, and transfer it from the protein ligand affinity prediction task to the RNA ligand affinity prediction task through transfer learning.

The entire method is divided into two stages:

- Pre training stage: Pre train the GatorAffinity model on the large-scale protein ligand structure affinity datasets SAIR and GatorAffinity-DB, allowing the model to first learn general atomic level 3D geometric and physical interaction rules (in this step, we use the officially trained weights and do not retrain).

- Migration fine-tuning stage: While keeping the model structure completely unchanged, replace the input from "protein-ligand" to "RNA-ligand", load pre training weights, fine tune on the RNA ligand subset of PDBbind, and perform 5-fold cross validation. We refer to the fine tuned model as GatorAffinity-RNA.

In terms of model architecture, we completely adopted the original design of GatorAffinity, but replaced the input from "protein ligand" to "RNA ligand" and changed the training data. For each

RNA ligand complex, we extract the atomic type (while labeling whether the atom belongs to RNA or ligand), predefined block types (such as bases in RNA and different functional groups in ligands), and the 3D coordinates of each atom from the structural file. Subsequently, the composition is divided into two scales: at the atomic scale, all atoms are treated as nodes, and edges are connected based on chemical bonds and spatial proximity relationships. The edges can carry features such as bond types and distances; At the block scale, atoms within the same base or fragment are aggregated into a block node, and a coarser grained graph is constructed using the centroid coordinates and type features of the block. In this way, the model can see fine atomic level interactions and perceive larger scale structural organization patterns.

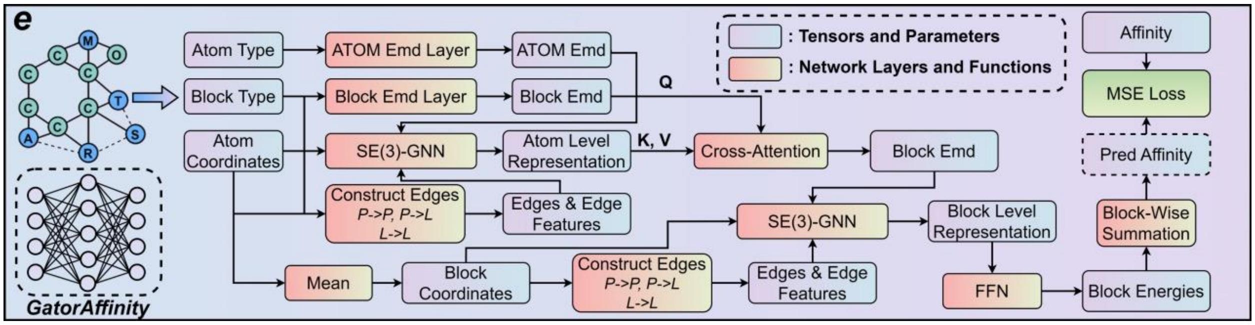


Figure 2 Model architecture of GatorAffinity and GatorAffinity-RNA

GatorAffinity uses SE (3) - equivariant graph neural network as its geometric modeling core: firstly, atomic types are mapped into vectors through embedding layers and sent to atomic level SE (3) - GNN together with 3D coordinates. Through multi-layer equivariant convolution, atomic representations encoding both chemical types and local spatial neighborhoods are obtained; These atomic features are aggregated within their respective blocks and added to the block type embeddings to form the initial representation of the block, which is then updated on the block graph through a second SE (3) - GNN. There is a layer of cross attention interaction between atomic level and block level: block representation serves as a query, "pulling" key local interactions related to itself from atomic representation to achieve cross scale information fusion. Finally, the network predicts a local energy contribution for each block, sums up the total energy of the complex across all blocks, maps it to an affinity prediction value (e.g. - logKd) through a linear layer, and trains it with mean square error as the loss function.

In the experiments, we first performed a simple cleaning of the RNA ligand subset of PDBbind, removing a few samples with obvious structural errors, and ultimately retained about 140 complexes for fine-tuning and evaluation. Subsequently, 5-fold cross validation was used: all samples were randomly divided into 5 approximately sized folds, and in each fold experiment, 1 fold was selected as the test set (about 20%), while the remaining 4 folds were merged into the training and validation sets. About 10% of the samples were selected as the validation set, and the remaining 70% were used as the training set; We fine tuned the pre trained GatorAffinity RNA on the training set and used the RMSE metric of the validation set for early stopping and hyperparameter selection. Finally, we evaluated the model performance on the corresponding test set, with a five fold rotation as the test fold. We reported the average performance and standard deviation of five experiments. All comparison methods, including AutoDock Vina, RSAPred, RLaffinity, and the w/o pretraining version without pretraining, use the exact same data partitioning and evaluation process to ensure fairness in the comparison.

We use four common metrics to evaluate model performance:

- RMSE (Root Mean Squared Error): The primary metric, measuring the root mean square error between predicted affinity and the true label;
- MAE (Mean Absolute Error): Measures the mean absolute error, more intuitively reflecting the magnitude of numerical deviation;
- Pearson Correlation Coefficient: Measures the linear correlation between predicted and true values, focusing on whether the overall trend is consistent;
- Spearman Correlation Coefficient: Calculated at the ranking level, reflecting the model's reliability in ranking candidate molecules.

Since we directly used the pre-trained model parameters from GatorAffinity, we didn't need to pre-train on 1.5 million datasets; we only needed to fine-tune them. Therefore, we completed all the fine-tuning tasks on a personal computer with a 16-core CPU, 32GB RAM, and an RTX 3090 24GB SSD.

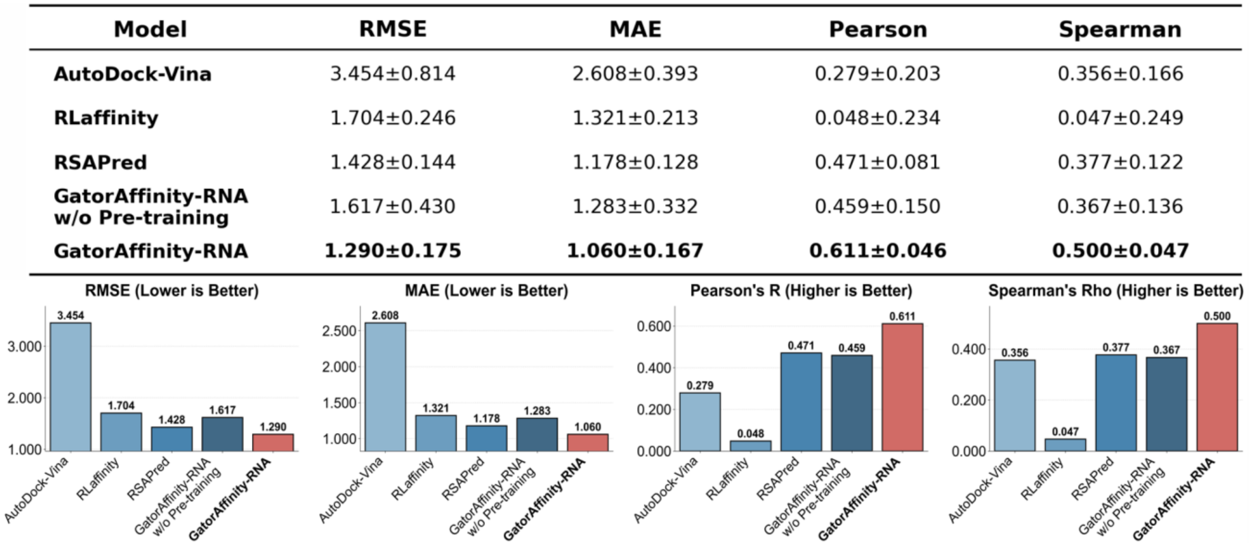
Results

We summarize the experimental results of GatorAffinity RNA on the PDBbind RNA dataset and compare them with several representative baseline methods. All results are based on five fold cross validation, with metrics including RMSE, MAE, Pearson correlation coefficient, and Spearman correlation coefficient.

Overall, GatorAffinity RNA achieved the best results in all four indicators after transfer learning. Under five fold cross validation, our model achieved RMSE 1.290 ± 0.175, MAE 1.060 ± 0.167, Pearson 0.611 ± 0.046, and Spearman 0.500 ± 0.047, outperforming all baseline methods and ablation versions without pre training in terms of numerical error and correlation.This result is also intuitively reflected in the bar chart visualization in the figure: GatorAffinity RNA has the lowest columns in the RMSE and MAE dimensions, and the highest columns in the Pearson and Spearman dimensions, showing an overall leading trend. From the training process perspective, we first attempt to directly train the SE (3) - GNN model ("w/o Pre training") from scratch on RNA data, which can be seen as an upper bound estimation that relies solely on small-scale RNA structures for learning. The RMSE of the model at a 5-fold is 1.617 ± 0.430, with a large error and significant standard deviation, indicating that training a geometric model from scratch on 140 sample level data is difficult to learn stable and generalizable physical interaction patterns. On this basis, we loaded GatorAffinity weights pre trained on 1.5 million protein ligand synthesis structure affinity data and fine tuned them on the same RNA data to obtain GatorAffinity RNA. After migration, the RMSE of the model decreased to 1.290 ± 0.175, and the Pearson correlation coefficient increased to 0.611 ± 0.046. Compared with the untrained version, the error was reduced by about 20.2%, and the correlation was significantly improved. This result suggests that atomic level geometric priors learned from protein ligand interactions can indeed be transferred to RNA ligand affinity tasks.

To verify that this improvement is not caused by random fluctuations between folds, we conducted paired t-tests on the five-fold results: the p-values for RMSE, MAE, Pearson, and

Spearman were 0.0177, 0.0134, 0.0081, and 0.0477, respectively, all significantly less than 0.05, indicating that the improvement brought by transfer learning is statistically significant.



Model	RMSE	MAE	Pearson	Spearman
AutoDock-Vina	3.454±0.814	2.608±0.393	0.279±0.203	0.356±0.166
RLaffinity	1.704±0.246	1.321±0.213	0.048±0.234	0.047±0.249
RSAPred	1.428±0.144	1.178±0.128	0.471±0.081	0.377±0.122
GatorAffinity-RNA w/o Pre-training	1.617±0.430	1.283±0.332	0.459±0.150	0.367±0.136
GatorAffinity-RNA	**1.290±0.175**	**1.060±0.167**	**0.611±0.046**	**0.500±0.047**

Figure 3 Summarizes the performance comparison of GatorAffinity-RNA with various baseline methods on PDBbind-RNA.

Regarding traditional docking methods, AutoDock-Vina achieved an RMSE of 3.454 ± 0.814 and an MAE of 2.608 ± 0.393, while Pearson's was only 0.279 ± 0.203. This indicates that empirical scoring functions exhibit significant errors and low correlation in RNA-ligand quantitative affinity prediction tasks, making them more suitable as pose search tools than high-precision affinity regressors. In deep learning baselines, RSAPred, as a sequence- and simple feature-driven model, achieved an RMSE of 1.428 ± 0.144, a MAE of 1.178 ± 0.128, and a Pearson score of 0.471 ± 0.081; while the structure-pre-trained model RLaffinity achieved an RMSE of 1.704 ± 0.246. This indicates that relying solely on one-dimensional sequence information or performing self-supervised pre-training on small-scale RNA structures is insufficient to fully capture the affinity-related 3D geometric and physical constraints.In contrast, our GatorAffinity-RNA further reduces RMSE and MAE by approximately 0.14 and 0.12 respectively compared to RSAPred, and also shows significant improvements in Pearson and Spearman scores (from 0.471/0.377 to 0.611/0.500), while significantly outperforming RLaffinity.

Discussion

The goal of this project is to test a specific hypothesis: whether the geometric priors learned from large-scale protein ligand synthesis structure affinity data can be utilized in RNA ligand affinity prediction to alleviate the performance bottleneck caused by the extreme scarcity of RNA structural data. To this end, we directly used GatorAffinity as the backbone model, pre trained it on approximately 1.5 million protein ligand synthesis complexes in SAIR and GatorAffinity DB, and then transferred the model to PDBind RNA for fine-tuning and evaluation, obtaining GatorAffinity RNA. We systematically compared it with traditional docking, existing deep learning methods, and "no pre training ablation".

From the results, the two-step "iteration" is very clear. The first step is to attempt to directly train the same SE (3) - equivariant GNN (w/o pre training) from scratch on RNA ligand small data without any protein pre training. Although the model structure is complex enough, it only achieved an RMSE of 1.617 ± 0.430 under five fold cross validation, with a large error and significantly high variance. This indicates that in scenarios with only about 140 samples, geometric models are difficult to "explore" stable physical interaction patterns on their own, and data scarcity makes the model prone to overfitting certain RNA families and unable to generalize. In the second step, we loaded the GatorAffinity weights trained on 1.5 million protein ligand synthesis structure affinity data and fine tuned the RNA data on identical partitions to obtain GatorAffinity RNA. After migration, the RMSE of the model decreased to 1.290 ± 0.175, and the Pearson correlation coefficient increased to 0.611 ± 0.046. Compared with w/o pre training, the error was reduced by about 20.2%, and the correlation was significantly improved; The p-values obtained from paired t-tests on all four indicators are less than 0.05, indicating that the performance improvement is not a random fluctuation, but indeed comes from the geometric and physical priors injected during protein ligand pre training.

There are still many limitations to this work. Firstly, the sample size and target diversity of the PDBbind RNA subset are limited, making it difficult to cover various types of RNA targets. Therefore, the current results are more like validating the effectiveness of the method within a narrow domain of "existing high-resolution structural and affinity data", rather than directly extrapolating to all RNA ligand scenarios. Secondly, the complex structures in SAIR and GatorAffinity DB are themselves synthetic structures predicted by the Boltz series models. Although they have been proven useful in protein tasks, their systematic biases and uncertainties may still be introduced into RNA models during migration. Again, our model only uses static 3D structures and does not explicitly consider the conformational dynamics of RNA, which may become a bottleneck when dealing with highly flexible RNA.

Based on these limitations, this work can be expanded in several directions in the future: firstly, attempting to construct RNA ligand synthesis structure affinity databases similar to SAIR/GatorAffinity DB, such as combining RNA structure prediction models and generative ligand modeling to provide richer pre training corpora for RNA ends; The second is to explore multi task or joint training methods, learning multiple tasks of protein ligand, RNA ligand, and even protein RNA complexes simultaneously in the same SE (3) architecture, further enhancing the generality of geometric priors

The code for our project can be found in https://github.com/wei00394-ship-it/CSCI-5526-25fall-Group4

Conclusion

The primary objective of this re-analysis was to investigate whether large-scale synthetic protein structure data could mitigate the critical challenge of data scarcity in RNA-ligand affinity prediction. Our results demonstrate that pre-training on protein structures generated by Boltz-1 significantly enhances model performance in RNA-ligand affinity prediction, enabling our model to achieve state-of-the-art accuracy (RMSE 1.290). These findings support the hypothesis that structural interaction patterns are transferable across biomolecular domains. In future work, we aim to explore generating synthetic RNA-ligand structures directly to further enrich the training domain. Progress in this direction may significantly improve the prediction accuracy of RNA-ligand affinity prediction models, thereby effectively reducing early-stage screening costs and accelerating drug discovery in RNA drug development.

References

[1] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, "The PDBbind database: methodologies and updates," *J. Med. Chem.*, vol. 48, no. 12, pp. 4111–4119, 2005.

[2] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, 2016.

[3] A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1100–D1107, 2012.

[4] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

[5] S. R. Krishnan, A. Roy, and M. M. Gromiha, "Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning," *Brief. Bioinform.*, vol. 25, no. 2, p. bbae002, 2024.

[6] S. Sun and L. Gao, "Contrastive pre-training and 3D convolution neural network for RNA and small molecule binding affinity prediction," *Bioinformatics*, vol. 40, no. 4, p. btae155, 2024.

[7] C. Pascual-Morena *et al.*, "Efficacy of risdiplam in spinal muscular atrophy: A systematic review and meta-analysis," *Pharmacother. J. Hum. Pharmacol. Drug Ther.*, vol. 44, no. 1, pp. 97–105, 2024.

[8] J. Wei *et al.*, "GatorAffinity: Boosting Protein-Ligand Binding Affinity Prediction with Large-Scale Synthetic Structural Data," *bioRxiv*, pp. 2025–09, 2025.

[9] C. Zhang, X. Li, Q. Guo, and S. Wang, "Sair: Learning semantic-aware implicit representation," in *European Conference on Computer Vision*, Springer, 2024, pp. 319–335.