Cross-Domain Multimodal Contrastive Learning for RNA-Ligand Binding Affinity Prediction

Tzu-Tang Lin
tzutang.lin@ufl.edu

Abstract

Expanding drug targets to RNA offers vast therapeutic potential, but RNA-ligand binding prediction remains underexplored due to limited datasets and the limitations of traditional methods. Conventional approaches, such as scoring functions and molecular docking, struggle to account for the complexity of RNA structures, solvent effects, and binding flexibility. Recent advancements in deep learning have demonstrated the ability to overcome these challenges by leveraging large-scale data and capturing intricate patterns in molecular interactions. This research introduces a multimodal contrastive learning model that combines interaction-free RNA and ligand sequence representations with interaction-based RNA-ligand binding pocket structural information. Inspired by the success of protein language models, the proposed model integrates structural knowledge into RNA sequences and small-molecule SMILES representations, creating a comprehensive cross-domain embedding. The pretrained foundation model is further fine-tuned using a transformer-based neural network to predict RNA-ligand binding affinity. The adoption of deep learning enables the model to handle the complexities of RNA-ligand interactions, capture non-linear relationships, and generalize effectively despite limited RNA-ligand datasets. Evaluation results demonstrate that the proposed model significantly outperforms existing methods, including retrained RLaffinity and transfer learning with GIGN. This study establishes a new benchmark in RNA-targeted drug discovery, showcasing the transformative potential of deep learning to address critical challenges in the field.

Contrbutions

1. Preprocessed and implemented a structure-based split of the PDBbind and Hariboss databases, creating high-quality datasets optimized for RNA-ligand binding affinity model development.
2. Retrained RLaffinity, the first deep learning-based method for predicting RNA-small molecule binding affinity, to improve performance on updated datasets.
3. Applied transfer learning using GIGN, a state-of-the-art model for protein-ligand binding affinity prediction, to explore its potential for RNA-ligand interactions.
4. Developed a cross-domain multimodal contrastive learning model for RNA-ligand binding affinity prediction

Introduction

Traditional small-molecule drugs focus on proteins, but only 10–15% of the 20,000 human proteins are disease-linked and "druggable," targeting just 0.05% of the genome. Expanding to RNA as drug targets, including mRNAs and non-coding RNAs, could unlock vast therapeutic potential. RNA comprises 70% of the genome, influencing disease pathways and enabling modulation of "undruggable" proteins. This shift offers a transformative opportunity to address unmet medical needs.[1] Recent studies have established that both coding and non-coding RNAs play significant roles in various diseases, positioning them as promising therapeutic targets.[2] As resistance to existing drugs increases and the pool of druggable proteins diminishes, especially in complex conditions like cancer and neurological disorders[3], RNA emerges as a critical focus for novel treatments. Advancements in RNA structural analysis have propelled this exploration, leading to the development of RNA-targeting therapies. A notable example is risdiplam (Evrysdi), an FDA-approved small-molecule drug that modifies RNA splicing to treat spinal muscular atrophy by increasing the production of functional survival motor neuron (SMN) protein.[4]

Binding affinity measures how strongly a drug interacts with its target, much like a key fitting into a lock.[5] These interactions occur at specific pocket-like regions of the target, known as binding sites. Stronger binding generally correlates with greater drug efficacy. Experimentally measuring binding affinity often relies on parameters such as the dissociation constant (Kd) and inhibition constant (Ki). However, these methods can be both costly and time-intensive, highlighting the importance of predictive approaches in drug discovery. This work focuses on developing a structure-based deep learning method to predict RNA-ligand binding affinity, offering a scalable and efficient solution for accelerating RNA-targeted therapeutic development.

Binding affinity prediction traditionally relied on scoring functions that estimate how well a ligand binds to a target[6]. Physics-based methods calculate binding energy using van der Waals, electrostatic, hydrogen bonding, and desolvation terms. Empirical methods, like Rosetta[7] and AutoDock[8], incorporate additional factors such as hydrophobicity, metal-ligand interactions, and entropy effects. Knowledge-based approaches learn from protein-ligand complexes to determine atom-pair potentials. Molecular docking simulations, using tools like AutoDock and Dock6[9], assess binding by predicting poses and assigning scores. While originally protein-focused, some methods were adapted for RNA-ligand docking. For example, AnnapuRNA evaluates RNA-ligand complex structures generated through docking[10]. Though effective, these approaches are now complemented by machine learning models, offering improved accuracy and scalability.

Conventional scoring functions are becoming outdated due to their reliance on independent energy terms, inability to model non-linearities, and neglect of factors like solvent effects and protein flexibility. In contrast, machine learning-based methods are revolutionizing binding affinity prediction and are rapidly becoming the mainstream approach. While protein-ligand binding prediction has been extensively studied using these methods, RNA-ligand binding remains relatively underexplored. Notable efforts in RNA-ligand binding prediction include RSAPred, which uses machine learning to predict RNA-small molecule binding affinity[11], and RLaffinity, which employs contrastive pre-training and 3D convolutional neural networks for RNA-small molecule binding affinity prediction[12].These studies highlight the growing interest and potential of machine learning in RNA-targeted drug discovery.

Given the limited size of RNA-ligand datasets and the lack of extensive prior research, leveraging insights from machine learning-based protein-ligand binding predictions is a promising strategy. The potential similarities between RNA-ligand and protein-ligand interactions make transfer learning a valuable approach for improving RNA-ligand model accuracy and performance, even with limited data. Protein-ligand methods can be categorized into interaction-free and interaction-based models. Interaction-free models predict binding affinity without focusing on physical interactions, learning representations from protein and molecule data, such as SMILES strings, sequences, and graphs. Notable examples include DeepDTA, which uses sequence-based representations[13] and GraphDTA, which combines protein sequences with molecular graphs[14]. Interaction-based models rely on the 3D structures of complexes, focusing on interactions at binding pockets, using 3D voxel grids or graphs for prediction. Examples include TANKBind, employing graph-based representations[15], and DeepAtom, using voxel-based 3D CNNs[16]. Adapting these approaches for RNA-ligand interactions, particularly using transfer learning, could bridge gaps in dataset size and significantly enhance prediction accuracy.

This research introduces a model that combines interaction-free RNA and ligand sequence representations with interaction-based RNA-ligand binding pocket structural information to create a multimodal contrastive learning pretraining foundation model. Inspired by recent advancements in protein language models, such as SaProt[17] and ESM3[18], which successfully incorporate structural knowledge, this model integrates pocket structure data into cross-domain RNA sequences and small-molecule SMILES representations. The pretrained foundation model is fine-tuned using its embeddings to train a transformer-based neural network for affinity prediction. Evaluation results indicate that this approach outperforms existing methods, including retrained RLaffinity and transfer learning with Geometric Interaction Graph Neural Network (GIGN) protein-ligand binding affinity prediction model[19], demonstrating enhanced performance in RNA-ligand binding affinity prediction.

Methods

1. Dataset Preparation

The nucleic acid-ligand (NL) dataset from PDBbind v2020 database[20], containing 149 complexes with binding data, was processed to extract RNA-ligand interactions. First, all complexes were downloaded from the PDB database using their PDB IDs. RNA structures were extracted using the RNA3DB method to transform non-standard residues[21], yielding 123 RNA complexes after excluding DNA-ligand interactions. Ligands were extracted based on provided ligand IDs, though six errors were identified (one DNA ligand, four protein ligands, and one missing ligand ID). Binding pockets were identified using a 6Å distance from ligands, resulting in 118 RNA-ligand interactions. To prevent bias and avoid overestimating model performance, RNA structures were clustered using the RMalign 3D structural alignment method[22]. Clustering with an RMscore threshold of 0.5 resulted in 41 clusters, with 22 unique groups combined into a single group labeled '0.' This process finalized the dataset into 20 groups, with a 7:1:2 train-validation-test split within each fold, ensuring the challenging group '0' was used for testing.

The HARIBOSS database[23], comprising 862 RNA-small molecule complexes, was similarly processed. Pocket records were validated and revised against PDB database entries, and non-

pocket-binding complexes, such as primers, were excluded. The same pipeline as PDBbind processing was applied: downloading all complexes, extracting RNA using RNA3DB methods, extracting ligands, and identifying pockets using the 6Å distance method. This processing resulted in 1,390 RNA-ligand interactions, which were used for pretraining.

2. GIGN transfer learning

The Geometric Interaction Graph Neural Network (GIGN)[19] is a state-of-the-art model for predicting protein-ligand binding affinities, leveraging 3D structural data and physical interactions. GIGN employs a heterogeneous interaction layer to independently process covalent and noncovalent interactions during the message-passing phase, ensuring effective node representation learning, and enforces translational and rotational invariance, making it robust to structural variations. To adapt GIGN for RNA-ligand binding prediction, transfer learning was applied, enabling the pre-trained protein-ligand binding model to fine-tune its knowledge on the smaller RNA-ligand dataset, thereby enhancing generalization and prediction accuracy. The RNA-ligand dataset was processed using GIGN's pipeline: RNA PDB and ligand MOL2 files were used to define binding pockets within 5Å of the ligand, which were then combined with the ligand PDB to create RNA-ligand complexes as RDKit MOL objects. These complexes were transformed into graph representations in PyTorch Geometric format for model training. Finally, GIGN was fine-tuned using the structure-based split of the dataset.

3. RLaffinity retraining

RLaffinity[12] is a pioneering deep learning method designed to predict RNA–small molecule binding affinity using 3D structural data. The model combines information from RNA binding pockets and small molecules through a 3D convolutional neural network (3D-CNN) and a contrastive learning-based self-supervised pre-training approach. In the pre-training phase, 1,415 RNA–small molecule complexes with interatomic distances smaller than 4Å were curated from the Protein Data Bank (PDB). For supervised training, 144 nucleotide –ligand pairs with binding affinity labels were obtained from the PDBbind database after filtering out DNA and peptide ligands.

Despite its novelty, RLaffinity has some limitations. Its use of a normalized negative log function (-logKd/Ki) scaled between 0 and 1 through min-max normalization deviates from the standard negative log function used in protein-ligand binding affinity predictions, potentially impacting consistency. Additionally, the PDBbind NL dataset used by RLaffinity includes both RNA and DNA-ligand complexes rather than RNA-ligand complexes only. To address these drawbacks, I retrained RLaffinity with the processed 118 PDBbind RNA-ligand dataset, ensuring RNA-ligand-specific predictions. The retraining process involved aligning the binding affinity representation with the standard negative log function (-logKd/Ki) and utilizing the same preprocessed RNA-ligand structure-based split datasets employed in GIGN transfer learning.

4. New model development

To address the challenges of RNA-ligand binding affinity prediction, the new model development focused on three key aspects. First, the issue of limited RNA-ligand datasets was tackled by pre-training on larger datasets to improve performance and generalizability. Second, a cross-modal approach was implemented to integrate RNA and ligand sequence representations with structural information, enabling more accurate predictions of RNA-ligand interactions. Finally, structural

knowledge was incorporated into cross-domain large language models (LLMs), leveraging their sequence-processing capabilities while enriching predictions with structural insights to capture the complexity of RNA-ligand binding.

Figure1 illustrates the model pretraining architecture, which integrates structural and sequence-based representations to model RNA-ligand interactions. For structural representation, RNA and ligand structures are merged into a pairwise co-distance matrix[24]. This matrix is constructed by extracting 3D atomic coordinates from RNA binding pockets and ligands, calculating pairwise Euclidean distances between RNA and ligand atoms, and populating a co-distance matrix where each entry represents the distance between specific RNA and ligand atoms. To ensure consistency across samples, the matrices are padded to a uniform size of 621×621.

For sequence-based representations, RNA sequences are embedded using CLS token embeddings from RiNALMo, the biggest RNA language model to date with 650M parameters, pre-trained on 36 million non-coding RNA sequences, producing embeddings with a fixed size of 1280 dimensions[25]. Ligands are represented using molecular token embeddings derived from SMI-TED, a SMILES-based Transformer Encoder-Decoder pre-trained on 91 million SMILES samples from PubChem, resulting in embeddings with a fixed size of 768 dimensions[26]. These sequence embeddings are concatenated and paired with the structure-based co-distance matrices for pretraining.

The pretraining model architecture utilizes a Contrastive Language-Image Pre-Training (CLIP) framework adapted for multimodal data[27]. It consists of three main components: a Vector Encoder comprising three fully connected neural network blocks for processing sequence embeddings, a Matrix Encoder with four convolutional neural network blocks for structural data, and a Cross-Modal Transformer Encoder with four transformer-based neural network blocks for integrating sequence and structure representations[28]. The outputs from the Cross-Modal Transformer Encoder are used to compute vector and matrix embeddings, which are optimized using a contrastive loss function. This approach facilitates the alignment between paired RNA-ligand sequence and structural representations while ensuring that unpaired sequence and structural representations are pushed farther apart. This contrastive learning framework enhances the foundation model's ability to effectively capture multimodal interactions by maximizing similarity between related pairs and minimizing similarity between unrelated pairs.

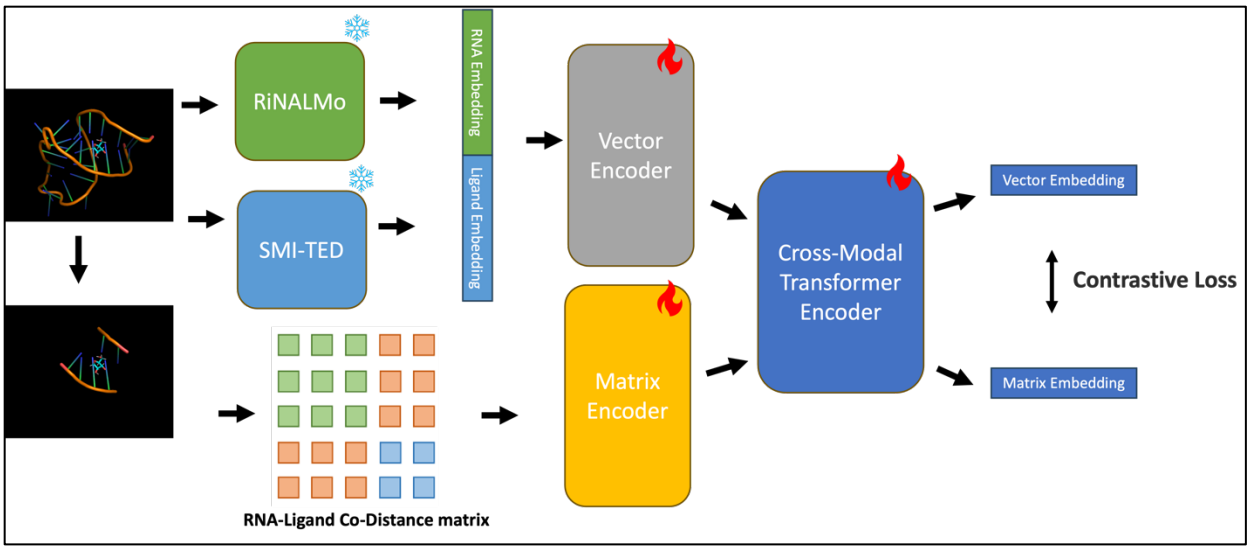

Figure1. Pretraining Model Architecture

Figure2 illustrates the fine-tuning approach for RNA-ligand binding affinity prediction, utilizing pretrained Vector Encoder, Matrix Encoder, and Cross-Modal Transformer Encoder with fixed weights to extract embeddings. These embeddings are used in a downstream transformer-based neural network comprising one transformer encoder layer and two fully connected layers. Three methods were employed to evaluate embedding contributions: Method 1 utilized only the Cross-Modal Transformer Encoder, Method 2 combined embeddings from the Cross-Modal Transformer Encoder, Vector Encoder, and Matrix Encoder, and Method 3 used embeddings from only the Vector Encoder and Matrix Encoder. Models were trained with three replicates under fixed random seeds, early stopping based on validation loss, learning rate scheduling, and Mean Squared Error (MSE) loss. This fine-tuning strategy assesses the effectiveness of different encoder combinations in capturing RNA-ligand binding affinity representations.

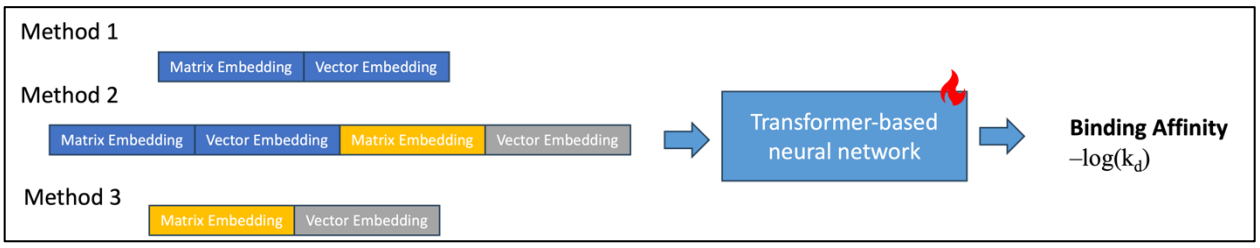

Figure2. Fine-Tuning settings for binding affinity prediction

Results

1. GIGN transfer learning and RLaffinity retraining

Table1 summarizes the performance of three approaches: RLaffinity Retraining, GIGN Retraining, and GIGN Transfer Learning. The RLaffinity Retraining model achieves a Test RMSE of approximately 1.9, outperforming GIGN Retraining (Test RMSE over 2.0) and demonstrating the effectiveness of using a contrastive learning-based RNA pretrained model for RNA-ligand binding affinity prediction. Similarly, GIGN Transfer Learning achieves a Test RMSE of about 1.9, also

outperforming GIGN Retraining and highlighting the benefit of leveraging protein-ligand binding models for RNA-ligand prediction. Overall, RLaffinity Retraining and GIGN Transfer Learning exhibit similar performance, with both achieving Test RMSE values of approximately 1.9, showcasing the promise of pretrained models for improving RNA-ligand binding affinity predictions.

Table1. Results of RLaffinity Retraining, GIGN Retraining, and GIGN Transfer Learning

Model	Replicate	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
RLaffinity Retraining	Rep1	1.5891	1.4572	0.2857834	**1.8717**	0.1338
RLaffinity Retraining	Rep2	1.5045	1.4442	0.2795134	**1.9133**	0.1969
RLaffinity Retraining	Rep3	1.5396	1.4762	0.2392457	**1.9556**	0.2192
GIGN Retraining	Rep1	1.0713	1.1823	0.6843	2.8779	0.1991
GIGN Retraining	Rep2	1.3303	1.0215	0.7474	2.1358	0.2299
GIGN Retraining	Rep3	1.1614	1.1006	0.7039	2.3292	0.1493
GIGN Transfer Learning	Rep1	0.7052	1.6345	0.053	**1.9585**	0.0992
GIGN Transfer Learning	Rep2	0.6197	1.5784	0.02	**1.8503**	0.1674
GIGN Transfer Learning	Rep3	0.8131	1.4151	0.5326	**1.9107**	0.0555

2. New model results

Table2 presents the performance of the newly developed model fine-tuned under three different settings, each utilizing different fixed-weight pretrained encoders to generate embeddings, followed by fine-tuning a transformer-based neural network for RNA-ligand binding affinity prediction. The results indicate that all three methods achieve similar performances, with a Test RMSE of approximately 1.8, outperforming the GIGN Transfer Learning and RLaffinity Retraining results (Test RMSE ~1.9) shown in **Table1**.

Among the three methods, Method 3, which uses only matrix and vector embeddings as representations, slightly outperforms Method 1 and Method 2, which include embeddings from the Cross-Modal Transformer Encoder. This suggests that matrix and vector embeddings alone are sufficient for effective binding affinity prediction, offering a marginal advantage over the inclusion of transformer embeddings. Overall, the results highlight the robustness and efficacy of the proposed model for RNA-ligand binding affinity prediction.

Table2. New Model Fine-Tuning Results

Method	Replicate	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
Method1	Rep1	0.706775	0.655467	0.922382	**1.859185**	0.122103

Method1	Rep2	0.653728	0.717947	0.929559	**1.798706**	0.087363
Method1	Rep3	0.742085	0.618312	0.935124	**1.782357**	0.12566
Method2	Rep1	0.800883	1.362149	0.484742	**1.759069**	-0.150561
Method2	Rep2	0.789578	1.3741	0.381237	**1.858708**	-0.114307
Method2	Rep3	0.811041	1.362216	0.407839	**1.889383**	-0.162566
Method3	Rep1	0.832534	1.480243	0.391878	**1.747118**	-0.114832
Method3	Rep2	0.815109	1.385777	0.434656	**1.795099**	-0.174513
Method3	Rep3	0.86197	1.479654	0.432712	**1.753488**	-0.117452

Discussion

1. Summary

In this research, I developed a cross-modal RNA-ligand binding affinity prediction model that integrates structural information into large language models (LLMs). By training a pretrained Contrastive Language-Image Pretraining (CLIP) model and fine-tuning it for binding affinity prediction using embeddings from the pretrained model, I demonstrated the model's capability for RNA-ligand interaction predictions. The proposed approach effectively combines sequence and structural information to achieve competitive performance, with results outperforming existing models like GIGN Transfer Learning and RLaffinity Retraining in terms of Test RMSE.

2. Future Work

Future work will focus on optimizing the model architecture and fine-tuning hyperparameters to enhance both performance and stability. While some early experiments showed promising performance with a Test RMSE of 1.68, the results were not consistently reproducible across repeated training runs. This highlights the potential of the model but also underscores the need for further refinement to ensure reliability. Additionally, the model's Pearson correlation (PR) performance remains suboptimal due to the current training loss focusing solely on minimizing RMSE. PR is a critical metric, especially for drug screening applications, where correlation between predicted and actual binding affinities is essential. Future efforts will involve incorporating additional loss functions to jointly optimize RMSE and PR, thereby improving the model's suitability for real-world drug discovery tasks.

References

1. Warner, K.D., Hajdin, C.E. & Weeks, K.M. Principles for targeting RNA with drug-like small molecules. *Nature Reviews Drug Discovery* **17**, 547-558 (2018).
2. Childs-Disney, J.L. et al. Targeting RNA structures with small molecules. *Nature Reviews Drug Discovery* **21**, 736-762 (2022).

3. Bernat, V. & Disney, M.D. RNA Structures as Mediators of Neurological Diseases and as Drug Targets. *Neuron* **87**, 28-46 (2015).
4. Sheridan, C. First small-molecule drug targeting RNA gains momentum. *Nature Biotechnology* **39**, 6-8 (2021).
5. Pantsar, T. & Poso, A. Binding affinity via docking: fact and fiction. *Molecules* **23**, 1899 (2018).
6. Liu, X. et al. Binding Affinity Prediction: From Conventional to Machine Learning-Based Approaches. *arXiv preprint arXiv:2410.00709* (2024).
7. Leman, J.K. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature methods* **17**, 665-680 (2020).
8. Huey, R., Morris, G.M. & Forli, S. Using AutoDock 4 and AutoDock vina with AutoDockTools: a tutorial. *The Scripps Research Institute Molecular Graphics Laboratory* **10550**, 1000 (2012).
9. Allen, W.J. et al. DOCK 6: Impact of new features and current docking performance. *Journal of computational chemistry* **36**, 1132-1156 (2015).
10. Stefaniak, F. & Bujnicki, J.M. AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses. *PLoS computational biology* **17**, e1008309 (2021).
11. Krishnan, S.R., Roy, A. & Gromiha, M.M. Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. *Briefings in Bioinformatics* **25**, bbae002 (2024).
12. Sun, S. & Gao, L. Contrastive pre-training and 3D convolution neural network for RNA and small molecule binding affinity prediction. *Bioinformatics* **40**, btae155 (2024).
13. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821-i829 (2018).
14. Nguyen, T. et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140-1147 (2021).
15. Lu, W. et al. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems* **35**, 7236-7249 (2022).
16. Li, Y., Rezaei, M.A., Li, C. & Li, X. in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 303-310 (IEEE, 2019).
17. Su, J. et al. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023.2010. 2001.560349 (2023).
18. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.2007. 2001.600583 (2024).
19. Yang, Z., Zhong, W., Lv, Q., Dong, T. & Yu-Chian Chen, C. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The journal of physical chemistry letters* **14**, 2020-2033 (2023).
20. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry* **48**, 4111-4119 (2005).
21. Szikszai, M. et al. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *bioRxiv*, 2024.2001.2030.578025 (2024).

22. Zheng, J., Xie, J., Hong, X. & Liu, S. RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC genomics* **20**, 1-10 (2019).
23. Panei, F.P., Torchet, R., Menager, H., Gkeka, P. & Bonomi, M. HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design. *Bioinformatics* **38**, 4185-4193 (2022).
24. Bryant, P., Kelkar, A., Guljas, A., Clementi, C. & Noé, F. Structure prediction of protein-ligand complexes from sequence information with Umol. *Nature Communications* **15**, 4536 (2024).
25. Penić, R.J., Vlašić, T., Huber, R.G., Wan, Y. & Šikić, M. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043* (2024).
26. Soares, E. et al. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267* (2024).
27. Radford, A. et al. in International conference on machine learning 8748-8763 (PMLR, 2021).
28. Lu, S., Gao, Z., He, D., Zhang, L. & Ke, G. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. *Nature Communications* **15**, 7104 (2024).