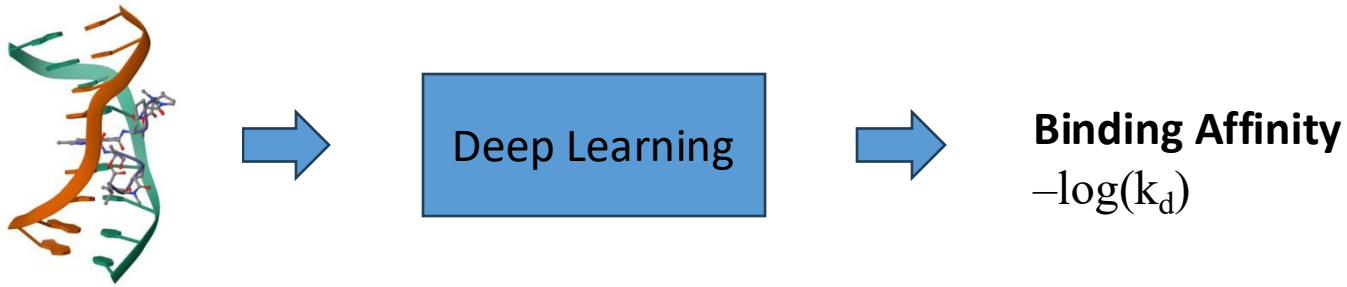
RNA-ligand Binding Affinity Prediction

Tzu-Tang Lin

Outline

1. Data preprocessing
2. Transfer learning --- Protein-Ligand binding affinity prediction model
3. Retrain existing --- RNA-Ligand binding affinity prediction model
4. New model development



PDBbind NL dataset process

- Downloaded **nucleic acid-ligand interactions** (NL) dataset from PDBbind v2020
- There are 149 nucleic acid-ligand complexes with binding data
- PDBbind NL dataset processing:

1. **Download all complexes** from PDB database using their PDB IDs.
2. **Extract RNA** using an RNA3DB method to transform non-standard residues.

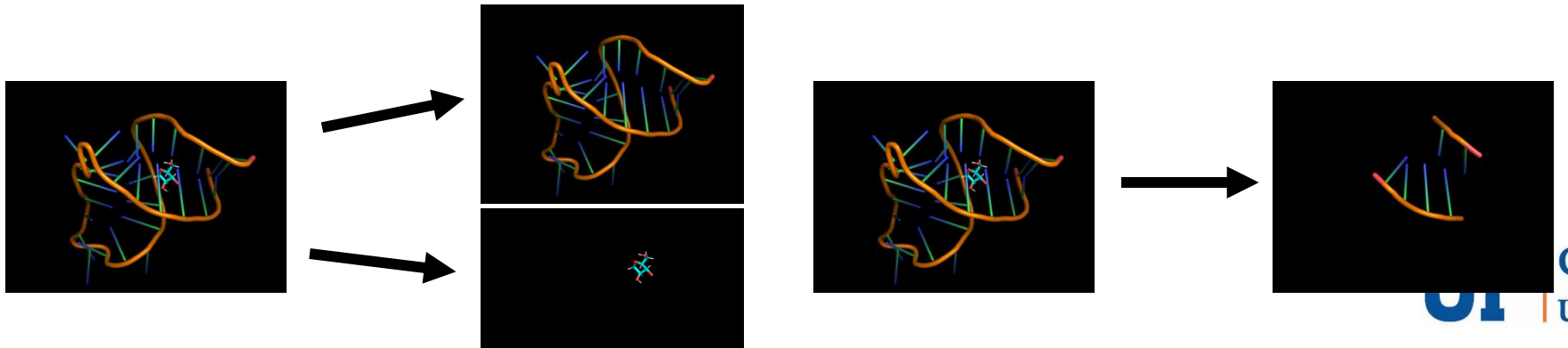
A total of 123 RNAs were extracted, indicating that the remaining complexes are DNA-ligand interactions.

3. **Extract Ligands** based on the given ligand IDs.

➢ 6 ligands were extracted with errors (1 was a DNA ligand, 4 were protein ligands, and 1 had a ligand ID that was not present in the complex.)

4. **Get Pockets** using a 6Å distance from ligands.

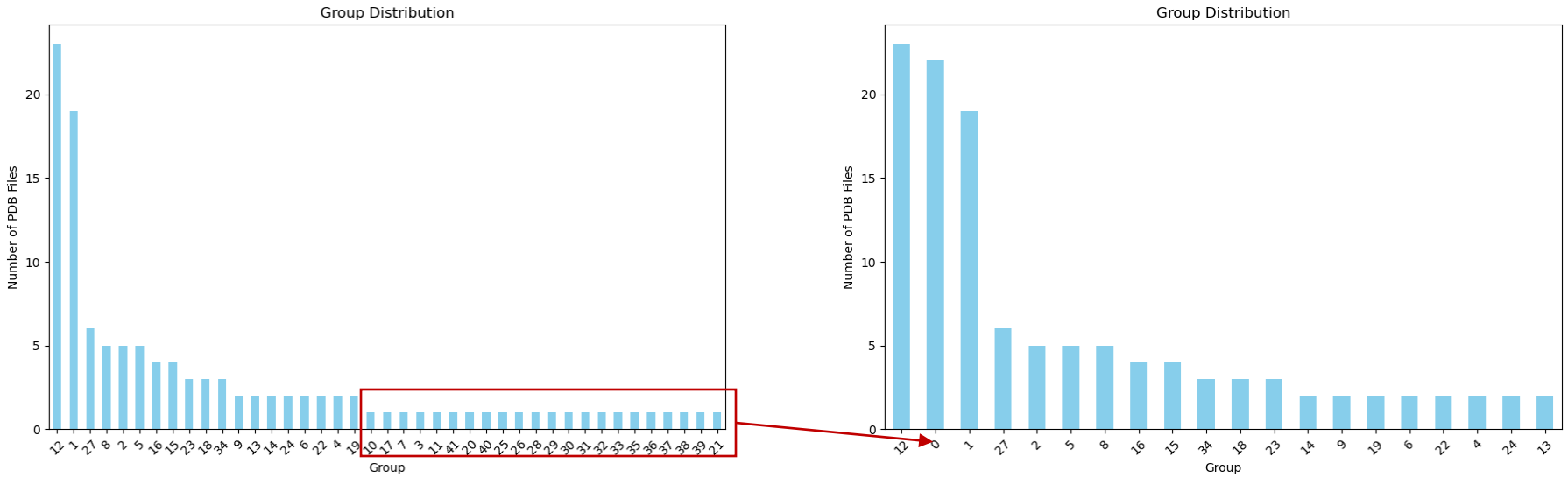
➢ **118** RNA-ligand interactions extracted (PDBbind RNA-ligand dataset)



PDBbind Structure-Based Split

- **Clustering by RNAs**:
  - Prevent bias and avoid overestimation of model performance using RNA 3D structural alignment (RMalign).
  - Grouped 118 RNA structures into 41 clusters using RMscore (threshold: 0.5). And then, combined 22 unique groups into a single group labeled ‘0’. This resulted in a final set of 20 groups for the dataset.
- **Split Strategy:**
  - Train-Validation-Test (7:1:2 ratio within each fold).
  - Most challenging group ('0') used as the test dataset.

➢ Generate the most challenging scenario for model development



UF College of Pharmacy UNIVERSITY of FLORIDA

GIGN transfer learning

- **Challenge:** Small RNA-ligand dataset size.

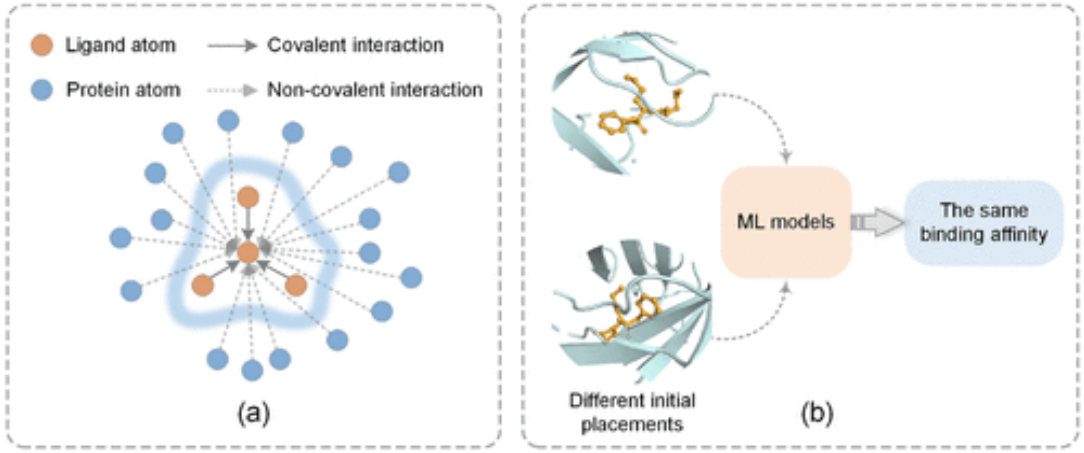
➢ Applied transfer learning using a protein-ligand binding model trained on a larger dataset.

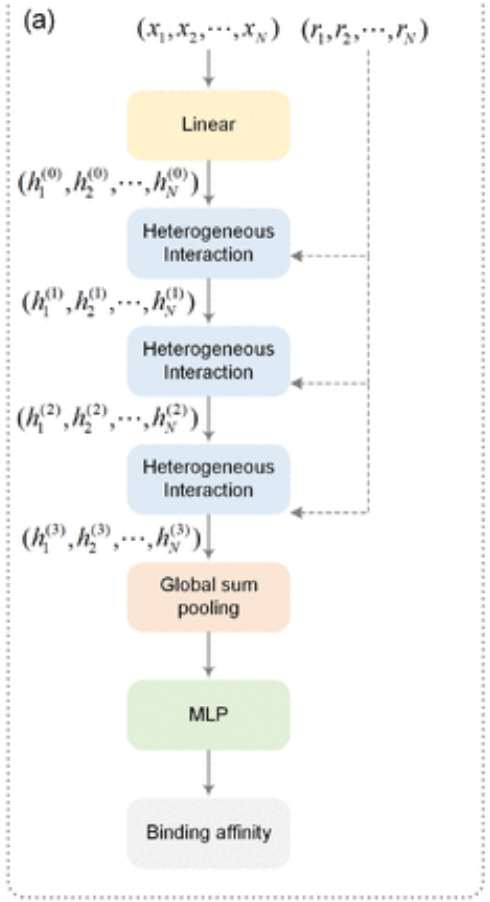
➢ Geometric Interaction Graph Neural Network (GIGN), a state-of-the-art model for protein-ligand binding affinity prediction.

GIGN Contribution:

1. Heterogeneous Interaction Layer: Processes covalent and noncovalent interactions independently.

2. Invariance Property: Ensures translation and rotation invariance in input data.





Geometric Interaction Graph Neural Network for Predicting Protein–Ligand Binding Affinities from 3D Structures (J. Phys. Chem. Lett., 2023).

rmacy
LORIDA

GIGN transfer learning

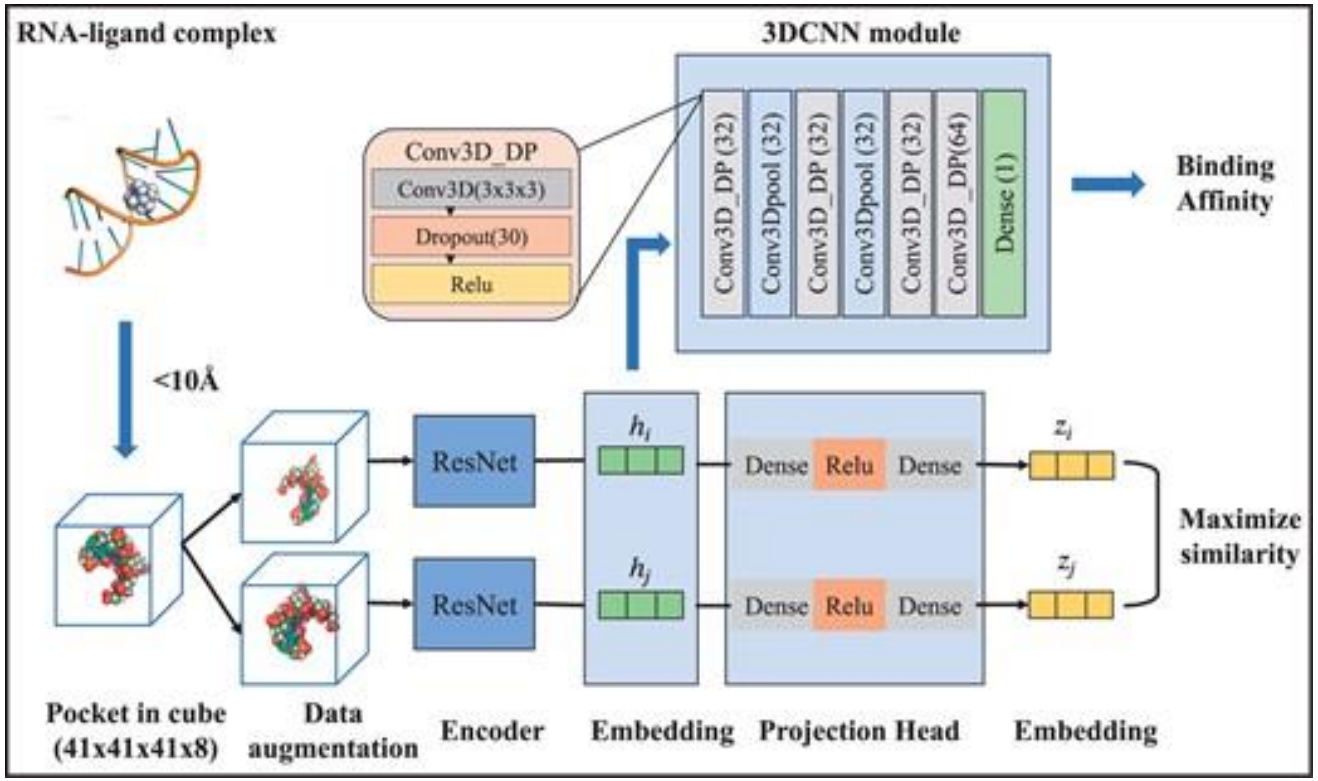
- Using GIGN data preprocessing for **118 RNA-ligand binding data**:
  1. **Get pockets**: Input the RNA PDB and ligand MOL2 files, define pockets within 5 Angstrom of the ligand (output: pocket PDB).
  2. **Generate complex**: Combine the pocket PDB and ligand PDB to create the complex (output: RDKit MOL).
  3. **Generate graph**: Use the complex (RDKit MOL) to generate a graph for model training (output: PyTorch tensor with PyG format).
- Fine-tuning and re-training based on GIGN's default parameters by structure-based split RNA-ligand binding dataset
- Each model trained with 3 replicates

Replicate	Training Phase	Train Loss	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
Rep1	Fine-tune	0.4973	0.7052	1.6345	0.053	**1.9585**	0.0992
Rep2	Fine-tune	0.384	0.6197	1.5784	0.02	**1.8503**	0.1674
Rep3	Fine-tune	0.6611	0.8131	1.4151	0.5326	**1.9107**	0.0555
Rep1	Re-train	1.1477	1.0713	1.1823	0.6843	2.8779	0.1991
Rep2	Re-train	1.7697	1.3303	1.0215	0.7474	2.1358	0.2299
Rep3	Re-train	1.3488	1.1614	1.1006	0.7039	2.3292	0.1493

➢ **Fine-tuning achieved lower test RMSE** compared to re-training, highlighting the advantage of leveraging protein-ligand binding models for RNA-ligand prediction

RLaffinity retraining

- RLaffinity was the first deep learning-based method for the prediction of RNA–small molecule binding affinity using 3D structures.
- RLaffinity integrated information from RNA pockets and small molecules, utilizing a 3D convolutional neural network (3D-CNN) coupled with a contrastive learning-based self-supervised pre-training model.



Contrastive Pre-Training and 3D Convolution Neural Network for RNA and Small Molecule Binding Affinity Prediction, *Bioinformatics*, April 2024

RLaffinity retraining

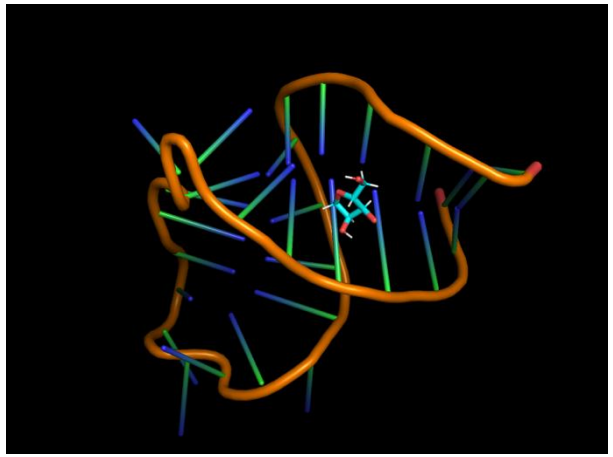
- Retrained RLaffinity models using the preprocessed structure-based Split RNA-ligand binding dataset same as GIGN transfer learning.

Replicate	Train RMSE	Train Pearson R	Val RMSE	Val Pearson R	Test RMSE	Test Pearson R
Rep1	1.5891	0.1191	1.4572	0.2857834	**1.8717**	0.1338
Rep2	1.5045	0.2735	1.4442	0.2795134	**1.9133**	0.1969
Rep3	1.5396	0.1887	1.4762	0.2392457	**1.9556**	0.2192

- The results are similar to the GIGN transfer learning (fine-tune) results

Replicate	Training Phase	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
Rep1	Fine-tune	0.7052	1.6345	0.053	**1.9585**	0.0992
Rep2	Fine-tune	0.6197	1.5784	0.02	**1.8503**	0.1674
Rep3	Fine-tune	0.8131	1.4151	0.5326	**1.9107**	0.0555

New model development



1. Data Limitations

- Pre-train on larger RNA-ligand datasets to improve performance.

2. Combine both sequence and structure information

- Develop cross-modal models for better RNA-ligand interaction predictions.

3. Combine structure information into Large Language Models

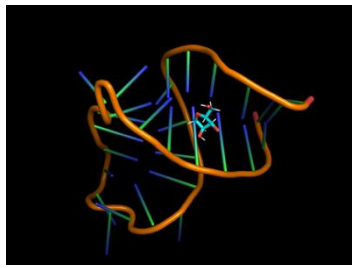
- Recently, SaProt (Su et al., 2024) and ESM3 (Hayes et al., 2024) have successfully added structural knowledge into protein language models

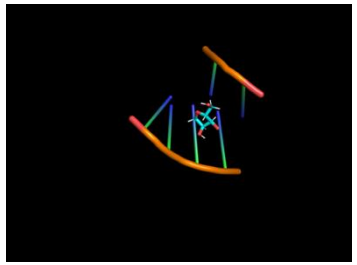
New model development – Hariboss Database

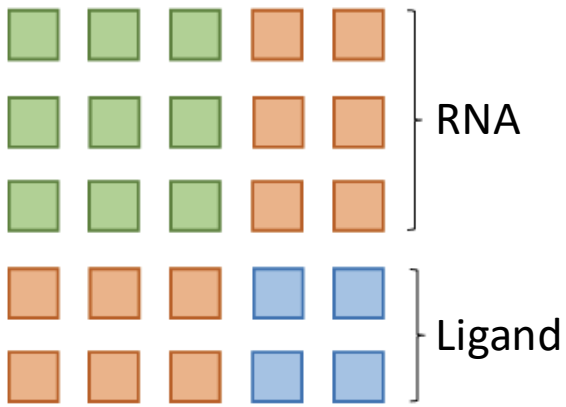
- HARIBOSS is a curated database of RNA-small molecules structures retrieved from the PDB
- HARIBOSS dataset processing:
  1. **Download** Got 862 RNA-SM complexes from the Hariboss database
  2. **Revise Error-Recorded Pockets** Compared pockets with PDB database entries.
  3. **Remove Non-RNA- Complexes** Excluded non-pocket binding complexes (e.g., primers).
  4. **Download all complexes** (same as PDBbind NL dataset processing)
  5. **Extract RNA** (same as PDBbind NL dataset processing)
  6. **Extract Ligands** (same as PDBbind NL dataset processing)
  7. **Get Pockets** (same as PDBbind NL dataset processing)

➢ **1390** RNA-ligand interactions extracted for pretraining

New model development – RNA-Ligand Structure representation





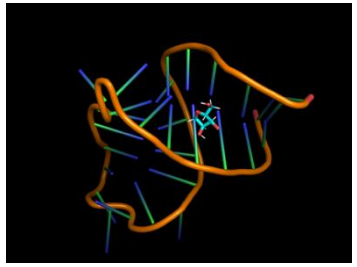


Pairwise Co-Distance Matrix

- **RNA-Ligand Complex**: Represent the RNA and ligand structure separately.
- **Steps to compute:**

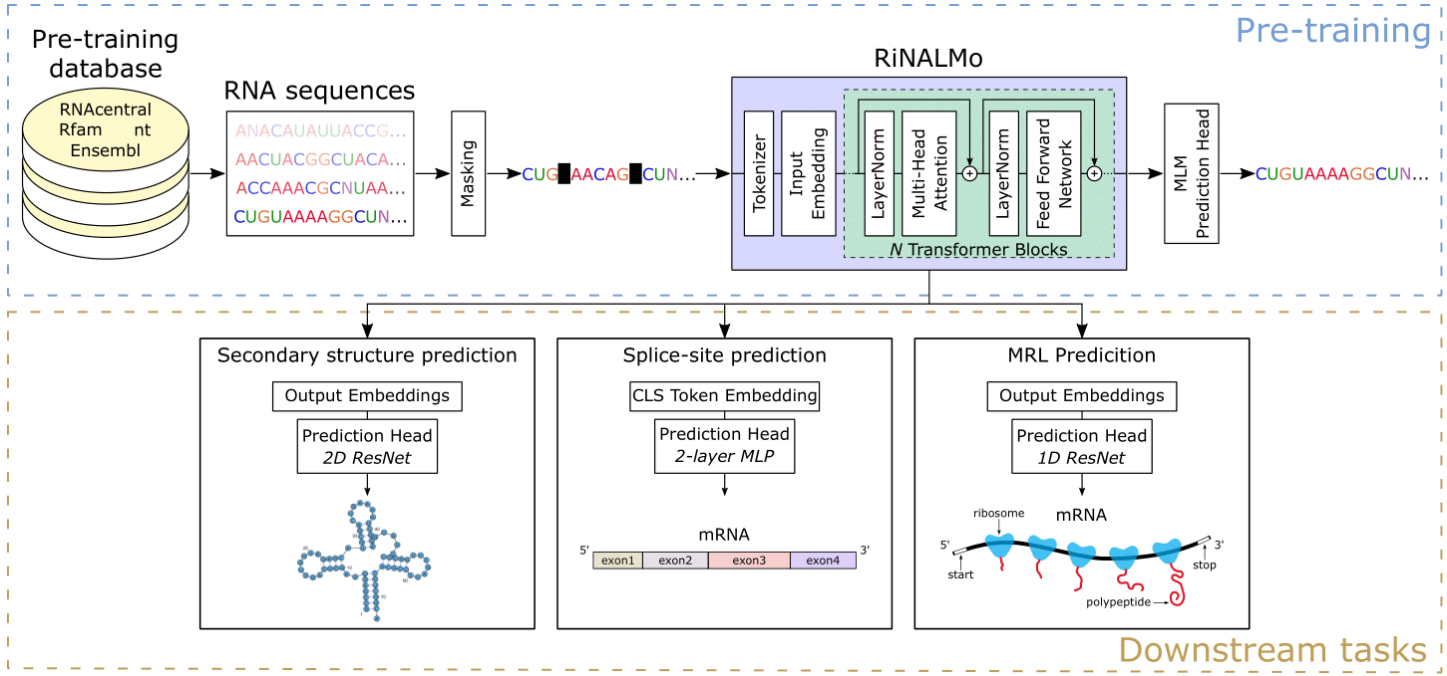
1. Extract 3D coordinates of atoms in the pockets and ligands.
2. Compute pairwise distances between all RNA and ligand atoms.
3. Construct a co-distance matrix where each entry represents the distance between a specific RNA atom and a ligand atom.
4. Pad all matrices to the maximum size in the dataset (Max Size: 621×621)

New model development – RNA sequence representation



RiNALMo

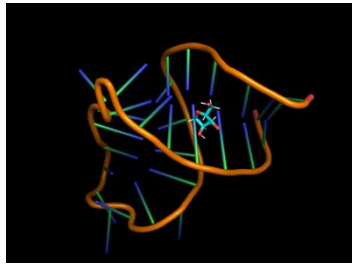
- RiNALMo is the largest RNA language model to date, with 650M parameters pre-trained on 36M non-coding RNA sequences from several databases

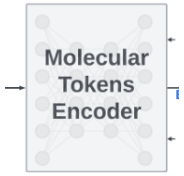


- Embedding size: 1280

RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks, arXiv, 2024

New model development – Ligands sequence representation

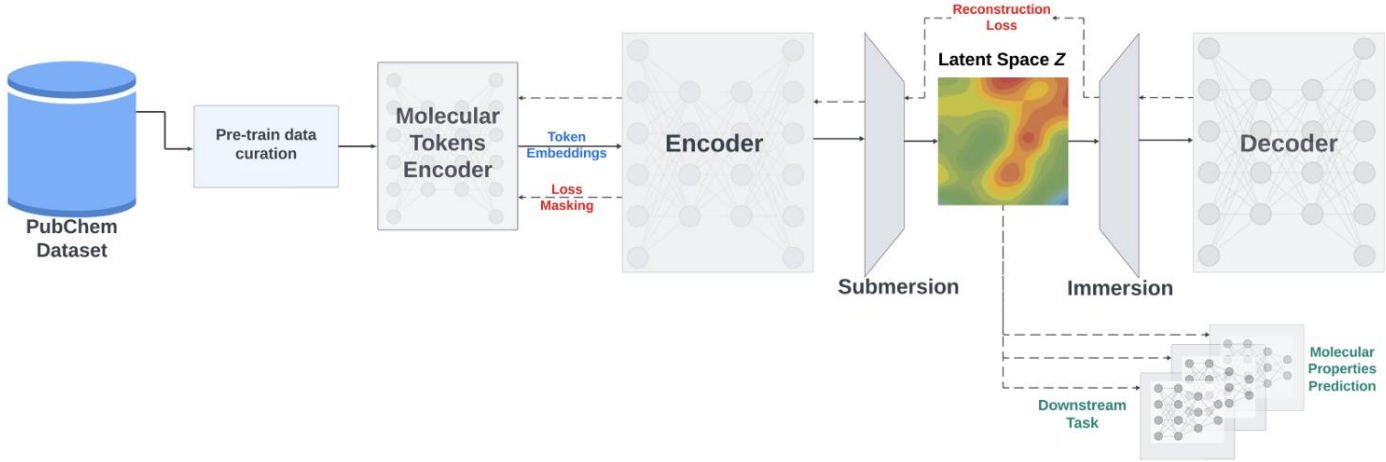




Token Embedding

SMI-TED

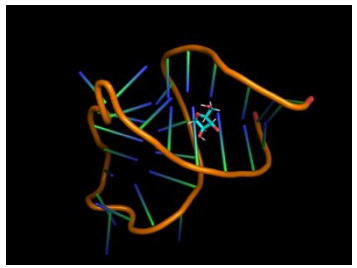
- SMILES-based Transformer Encoder-Decoder (SMI-TED), pre-trained on a curated dataset of 91 million SMILES samples sourced from PubChem, equivalent to 4 billion molecular tokens.

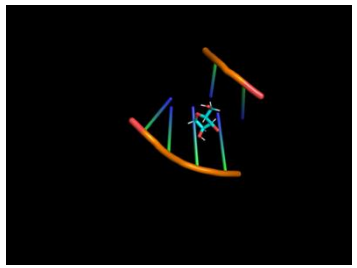


- Embedding size: 768

A Large Encoder-Decoder Family of Foundation Models For Chemical Language , arXiv, 2024

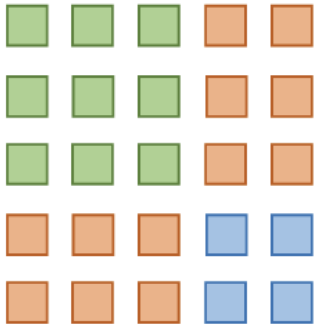
New model development – Contrastive Learning





RNA Embedding

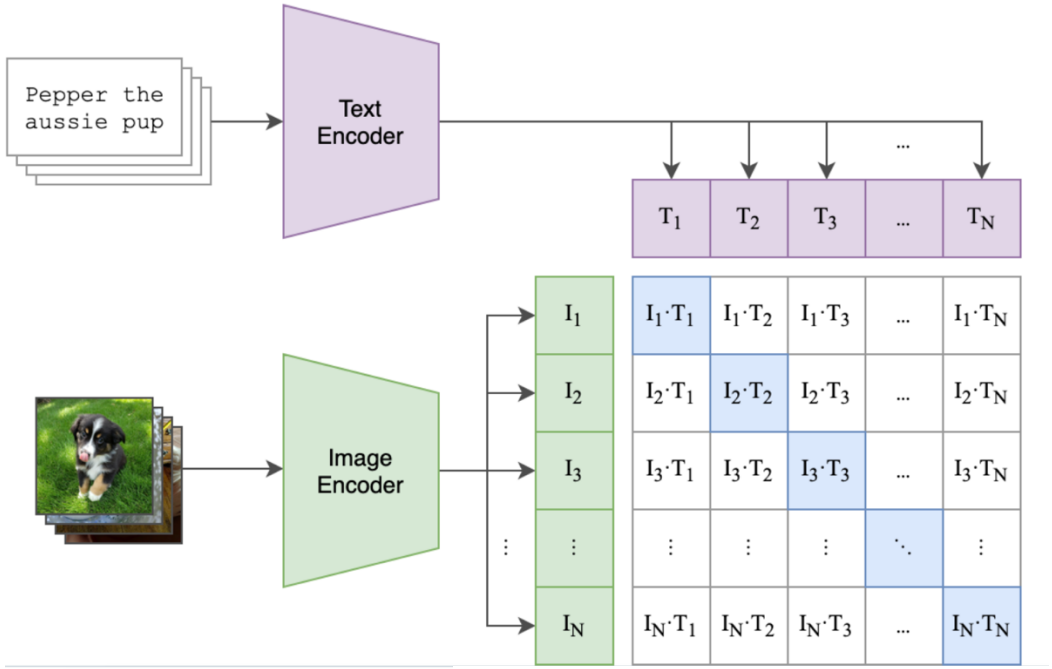
Ligand Embedding



RNA-Ligand Co-Distance matrix

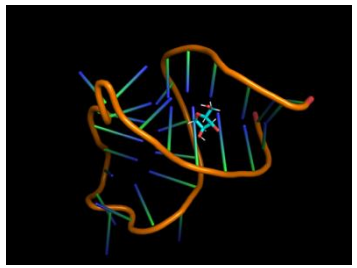
CLIP

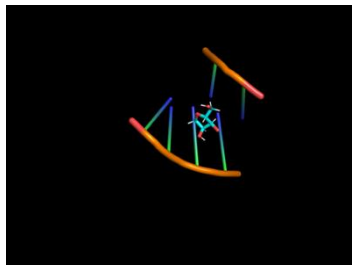
- CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of pairs.
- A multi-modality model for connecting texts and images




```
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

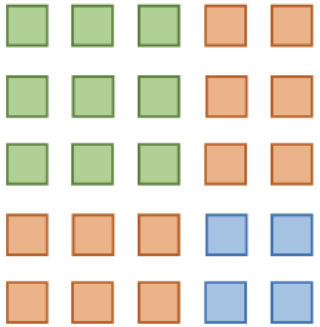
New model development – Contrastive Learning





RNA Embedding

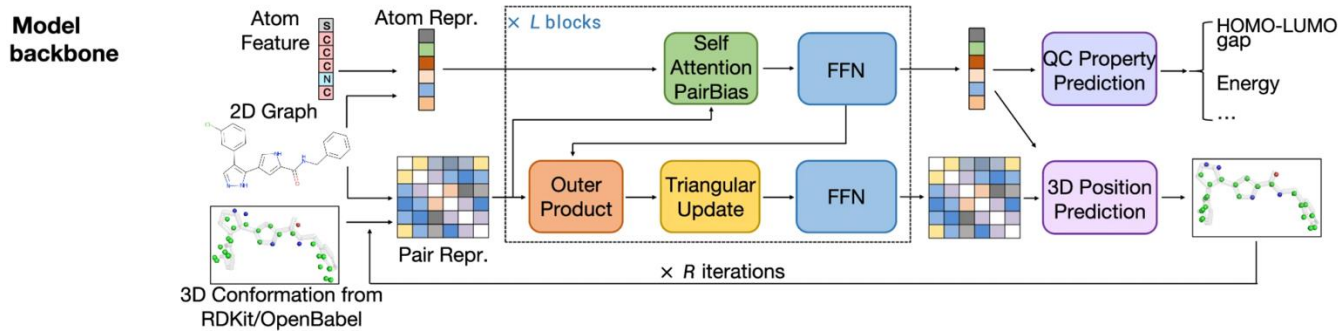
Ligand Embedding



RNA-Ligand Co-Distance matrix

Uni-Mol

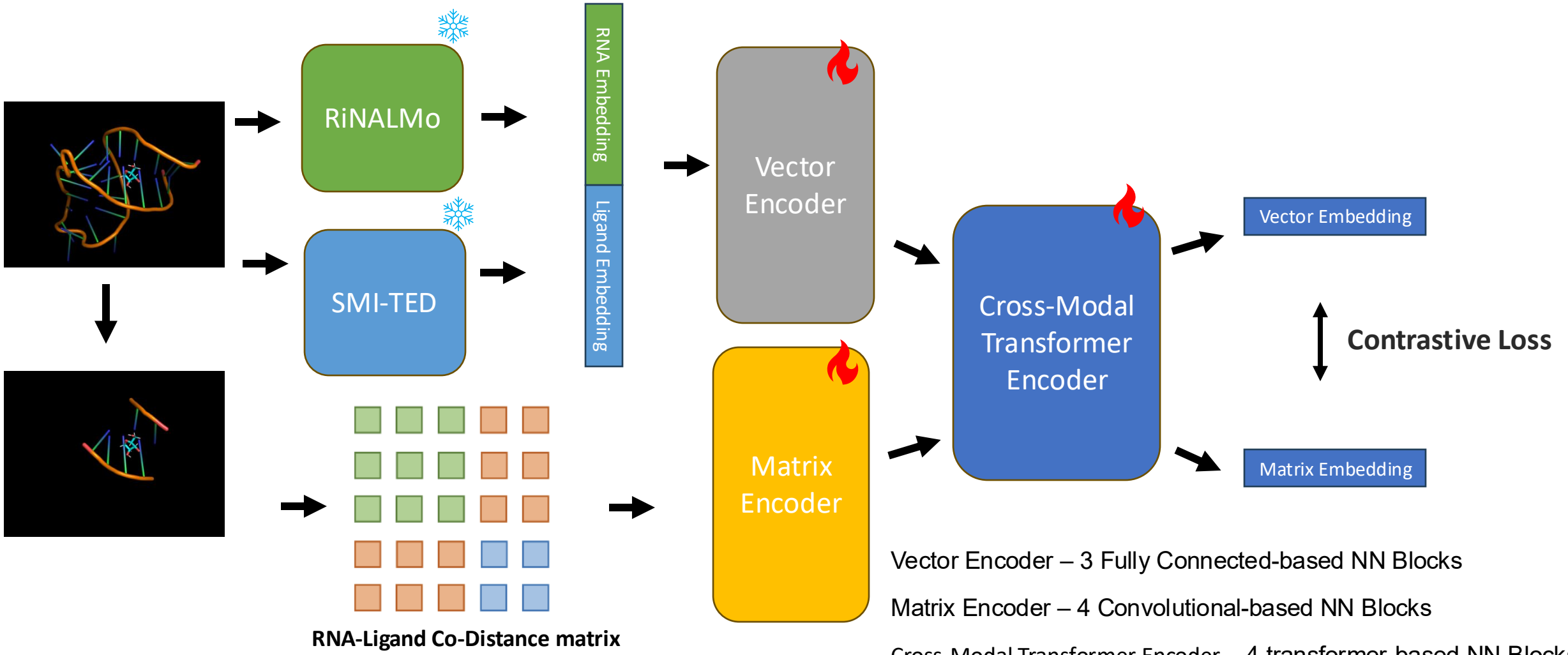
- The Uni-Mol modelling series described the **pretraining of general molecular encoders** and showcased their applications in various 2D and 3D downstream tasks.



- Got inspired by Uni-Mol backbone's TransformerEncoderWithPair modules, I developed the following pretrain model.

Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. Nat Commun, 2024

New model development – Pretrain

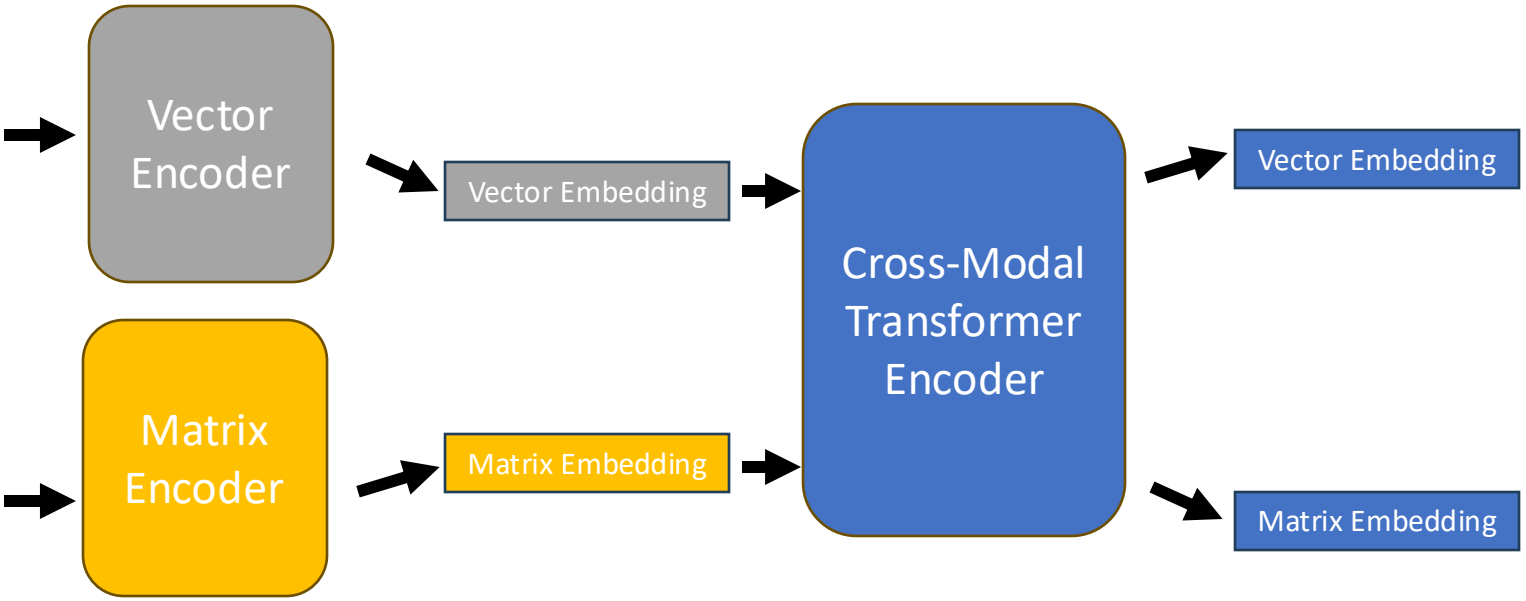


Vector Encoder – 3 Fully Connected-based NN Blocks

Matrix Encoder – 4 Convolutional-based NN Blocks

Cross-Modal Transformer Encoder – 4 transformer-based NN Blocks

New model development – Finetune



Finetune Strategy

- **Method 1**

Cross-Modal Transformer Encoder

- **Method 2**

Cross-Modal Transformer Encoder +

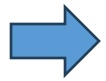
Vector encoder + Matrix encoder

- **Method 3**

Vector encoder + Matrix encoder

Method 1

Method 2



Binding Affinity

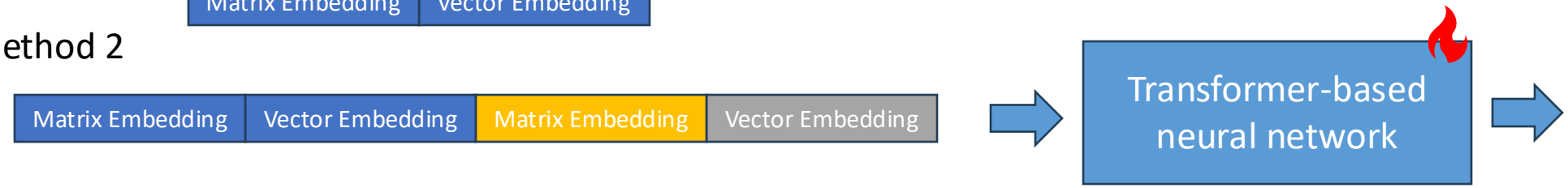
$-\log(k_d)$

Method 3

New model development – Results

Method 1

Method 2



Method 3

Method	Replicate	Train Loss	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
method1	Rep1	0.49953	0.706775	0.655467	0.922382	**1.859185**	0.122103
method1	Rep2	0.42736	0.653728	0.717947	0.929559	**1.798706**	0.087363
method1	Rep3	0.55069	0.742085	0.618312	0.935124	**1.782357**	0.12566
method2	Rep1	0.641414	0.800883	1.362149	0.484742	**1.759069**	-0.150561
method2	Rep2	0.623433	0.789578	1.3741	0.381237	**1.858708**	-0.114307
method2	Rep3	0.657787	0.811041	1.362216	0.407839	**1.889383**	-0.162566
method3	Rep1	0.693114	0.832534	1.480243	0.391878	**1.747118**	-0.114832
method3	Rep2	0.664402	0.815109	1.385777	0.434656	**1.795099**	-0.174513
method3	Rep3	0.742992	0.86197	1.479654	0.432712	**1.753488**	-0.117452

Training Setting

1. Fixed random seed
2. Early Stopping by Valid Loss
3. Learning Rate Scheduling
4. Mean Squared Error Loss
5. 1 transformer encoder layers, and 2 fully connected layers

- Three methods have similar performances
- Outperforms both GIGN transfer learning models and RLaffinity retrain models (Test RMSE: 1.8-1.9)

New model development – Discussion

Summary

- In this research, I developed a cross-modal RNA-ligand binding affinity prediction model
- By incorporating structural information into LLMs, I trained a pretrained CLIP model, then fine-tuned binding affinity prediction models using the pretrained model embedding

Future Work

- Future work will focus on **optimizing the model architecture and fine-tuning hyperparameters** to enhance performance and stability
- Some early experiments showed the model could perform well, but the results were not always consistent across repeated training runs. This shows the model has potential, but more work is needed to make it stable and reliable

Method	Type	Train Loss	Train RMSE	Valid RMSE	Valid PR	Test RMSE	Test PR
method1	Best	14.825213	3.850353	1.397021	0.663954	2.718086	0.117585
method1	Last	2.849923	1.688172	1.873621	0.206692	**1.682905**	0.183645
method2	Best	7.373564	2.715431	1.476381	0.508836	1.769341	-0.145751
method2	Last	3.301495	1.817002	4.741527	0.586770	3.996764	0.051575
method3	Best	6.091993	2.468196	1.471873	0.510478	**1.683729**	-0.203566
method3	Last	4.126519	2.031384	6.918242	0.471184	6.137150	-0.066232

Thank you!