Predicting Ligand – RNA Binding Using E3-Equivariant Network and Pretraining

Zhenfeng Deng[†‡]
DP Technology
Beijing, China
1810307204@pku.edu.cn

Ruichu Gu[†]
DP Technology
Beijing, China
gurc@dp.tech

Hangrui Bi[†]
DP Technology
Beijing, China
bihr@dp.tech

Hongli Ma
Tsinghua University
Beijing, China
hongli.ma.explore@gmail.com

Xinyan Wang
DP Technology
Beijing, China
wangxy@dp.tech

Zhaolei Zhang[*]
University of Toronto
Toronto, ON
zhaolei.zhang@utoronto.ca

Han Wen[*]
DP Technology
Beijing, China
wenh@dp.tech

Abstract

It is becoming increasingly appreciated that small molecules hold great promise in targeting therapeutically relevant RNAs, such as viral RNAs or splicing junctions. Yet predicting ligand targeting RNA is particularly difficult since limited data are available. To overcome this, we fine-tuned a pretrained small molecule representation model, Uni-Mol, to predict the RNA-binding propensity of ligands and the RNA binding QSAR model. In addition, we develop an E3-equivariant model to predict possible ligands given the RNA pocket geometry. To the best of our knowledge, this is the first E3-equivariant model for predicting RNA-ligand binding. We demonstrated the great potential of Uni-Mol pretraining in the RNA-ligand tasks towards efficient and rational RNA drug discovery.

1 Introduction

The human genome consists of approxiately 21,000 protein coding genes; however, currently only about 15% proteins [1, 2, 3] are deemed druggable by small molecule drugs. These small molecules typically bind to cavities or pockets in a protein molecule, thus either blocking the functional sites or blocking interaction between the protein and other partners. Unfortunately, not every protein has a "druggable" pocket. An alternative approach is to target the upstream RNA, either by anti-sense RNAs, aptamers or small molecules. Similar to proteins, RNA molecules also folded into structured entities in three-dimensional space, affording pockets or clefts for small molecules binding. [4]. Several successful lead compounds have been identified by high-throughput screening[5]. For example, Roche and PTC Therapeutics' risdiplam (an RNA splice-modifying small-molecule drug) has been approved for the treatment of spinal muscular atrophy (SMA)[6]. Motivated by these successes, several databases have been developed to curate experimentally determined small molecule and

[†]Equal Contribution
[*]Corresponding Author
[‡]School of Pharmaceutical Sciences, Peking University

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

RNA interactions, e.g., R-Bind[4], RNALigands[7], and HARIBOSS[8]. Several computational algorithms, including RBind, RNAmigos and others, were developed to learn the rules of small molecule and RNA interactions in order to facilitate in silico screening or design lead drug compounds [5]. In the following, we first review several representative methods and then introduce our proposed pretraining-based method.

Given these curated interaction data, there are several approaches in extracting information into predictable models, ranging from structure docking or similarity [9], similarities in RNA sequence and secondary structures (RNAligands), and more advanced machine learning approaches. These machine learning algorithms typically model the ligand pocket as a graph, with nucleotides represented as nodes and interaction between nucleotides or between nucleotides and ligands as edges in the graph. Features are assigned to these edges, depending on the type of interactions, i.e., Watson Crick hydrogen bonds or non-canonical hydrogen bonds [10]. Despite their advantages, these approaches may lose 3D information when processing structural data. Towards this goal, we herein described an improved method consist of E3-equivariant molecular representations given by a pretrained model, i.e. Uni-Mol (see Methods) for small molecules, and also an E3-equivariant description of RNA pockets. Our results firstly showed better performance in discriminating RNA binding and non-binding ligands, as well as RNA-focused QSAR. In addition, given 3D structures of RNA pockets, our RNA-ligand model can predict favorable chemical fingerprints or Uni-Mol representations for bound ligands.

2 Methods

We used a pretrained Uni-Mol network to embed small organic molecules in a latent space of 512 dimensions for both the ligand-based model and the RNA-based model. The whole scheme is described in Figure 1. For the ligand-based model, we trained a binary classifier to predict whether a ligand binds to RNAs given its Uni-Mol embedding and a QSAR model. For the RNA-based model, we employed an E3-equivariant regressive model to predict the chemical fingerprint or Uni-Mol embedding representation of favorable binding ligands, given the 3D structures of the RNA pockets.

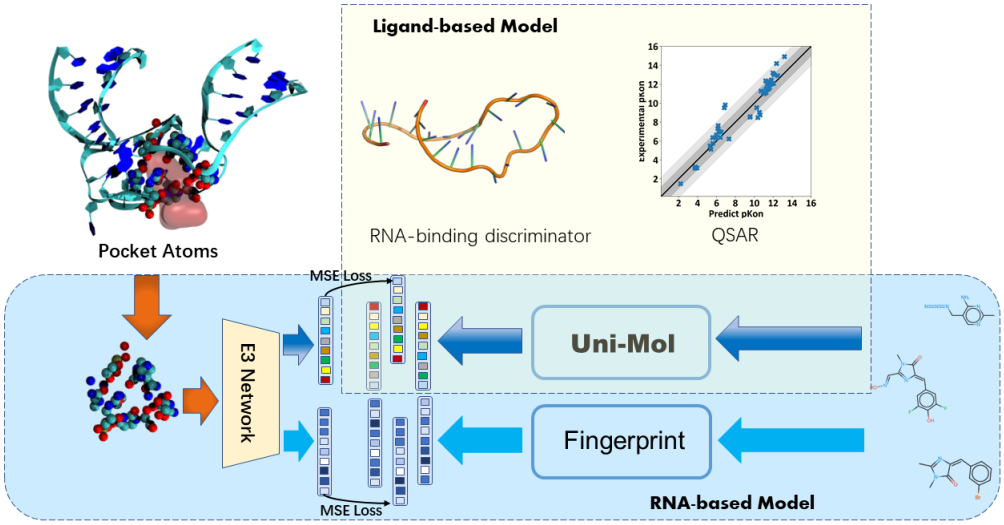

Figure 1: The Overall scheme of our model. Ligands pass through Uni-mol module to generate embedding vectors, for RNA-binding discriminator or QSAR use; can also generate fingerprint directly. Atoms of the RNA pockets are processed by E3 network to generate fingerprints or Uni-mol embedding vectors

2.1 Dataset Curation

The ligand based model contains two tasks, the classification, and the regression. For the classification task, we used the ROBIN database, the largest experimentally derived library of nucleic acid binders to

date.[5] The augmented set of ROBIN was used, which contains 2003 RNA-binding small molecules and 77678 protein-binding small molecules as the negative sets. Following the same protocol in ROBIN, the negative set and positive set are balanced by similarity search and repeated sampling. We extracted the smiles corresponding to the molecules in the augmented set and used RDKit[11] to generate the 3D conformations , then passed to the Uni-Mol[12] model for subsequent tasks.

For QSAR task,we processed the k_{on}, k_{off} and the binding affinity in Cai et. al.[13] with logarithmic transformation to be consistent with their protocol.

For pocket-based tasks, we retrieved all PDB entries from HARIBOSS[8], a database of RNA-small molecule structures. Ligands are extracted and converted to SMILES by OpenBabel[14] as the inputs. All heavy atoms within 4.5 angstroms of the ligands are extracted as the pockets. Only the first frame of NMR structures will be kept and low quality structures were removed. A total of 305 ligands and 970 pocket-ligand pairs were finally obtained and then split by a ratio of 810:90:70 for train/valid/test, respectively. For the downstream screening task, another set of 195 RNA targeting ligands from R-BIND2.0[4] was used instead.

2.2 Molecular Representation Generation

Uni-Mol is a molecule representation framework [12], using an E3-equivariant transformer architecture that pretrained on token prediction and geometric denoising on a dataset of 19M molecules, and achieved SOTA in 14/15 molecular property prediction tasks.

We take SMILES as inputs and re-generated 3D conformations following standard Uni-Mol protocol, namely the ETKGD [15] with Merck Molecular Force Field [16] optimization in RDKit[11]. For each entry it generates 10 3D conformations and a 2D conformation. After that, a 512-dim embedding for each conformation's [CLS] token (see [12] for details) is inferred with the molecular-pretrain model in Uni-Mol's official repository. The means vector of all conformations was used as the embedding of the input ligands.

MDL Molecular Access Keys (MACCS) fingerprint [17] is a 166-dim binary fingerprint of various chemical properties. It is widely used in property prediction and was adopted by a previous graph based screening model, RNAmigos[10]. We generate the MACCS fingerprints with OpenBabel[14] from the SMILES.

2.3 E3-Equivariant RNA Pocket Representation

We employed an E3 equivariant graph neural network[18, 19, 20] to predict the binding ligand given the RNA structure. The RNA structure is represented as the coordinates of each non-hydrogen atoms $\{x_i\}$, the atom type $\{Z_i\}$ and the set of chemical bond adjacency matrix $\{e_{ij}\}$. For each atom, its representation is a tensor $h_i^{l,p,m,c}$ for each degree $0 \leq l \leq l_{max}$, magnetic quantum number $-l \leq m \leq l$, parity $p \in \{odd, even\}$, and channel c. We initialize the invariant components $h_i^{l=0,even}$ with a trainable atom type embedding $embed_\theta(Z_i)$, and zero-initialize other components. We then updated the atom representations with multiple-layer message passing network, which is the backbone of our model. The message passed from atom a to atom b is (linear layers ommited for simplicity)

$$m_{a,b}^{l_o,p_o} = \sum C_{l_i,l_f}^{l_o} R_\theta^{l_f,p_f,l_i,p_i}(||x_a - x_b||, e_{ab}) \cdot h_i^{l_i,p_i} \cdot Y^{l_f}(x_a - x_b) \tag{1}$$

,where $C_{l_i,l_f}^{l_o}$ is the Clebsch–Gordan (CG) coefficients, $R_\theta^{l_f,p_f,l_i,p_i}$ is a neural-network-parameterized radial function, and Y^{l_f} is the spherical harmonics. For each step, we only considered the interaction between atom pairs between whom the distance is within 4 Å.

Since the molecular fingerprint and UniMol embedding are E3 invariant, only scalar components can be used to predict the ligand. Therefore we took the L2-norm of higher order tensors $\bigoplus_{l,p} \sum_m h_i^{l,m,p} * h_i^{l,m,p}$ and concatenate it with $h_i^{0,even}$ to obtain an invariant representation t_i for each atom. We then use a softmax weighted pooling of the invariant atom representations to get the final predicted ligand embedding t.

$$t = \Sigma_i \frac{e^{s_i}}{\Sigma e^{s_i}} t_i \tag{2}$$

,where s_i is a linear function of t_i. MSELoss head and Adam optimizer were used for training.

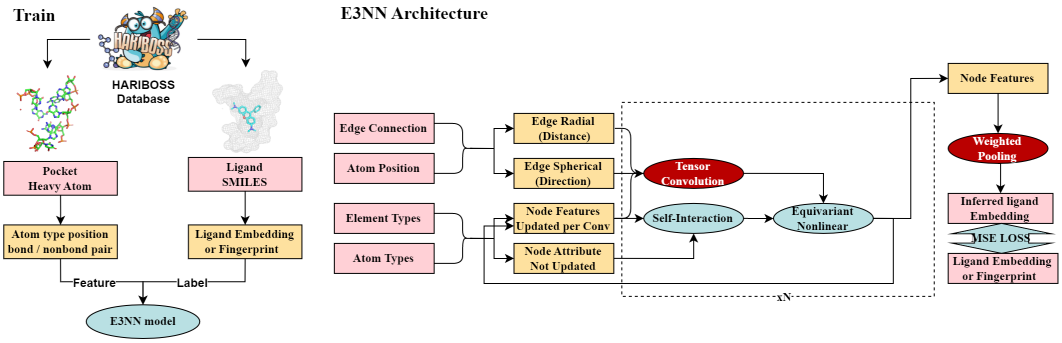

Figure 2: An overview of our E3-equivariant model. The atom features are initialized with atom type embeddings and updated by a stack of equivariant message passing modules. Readers may refer to Nequip[20] for the details of Self Interaction and Equivariant Nonlinearity.

2.4 Downstream Tasks

Two tasks were performed for the RNA based model (Figure 3). For the systematic retrieve task, we followed the screening protocol in RNAmigos[10]: for each RNA target pocket in the HARIBOSS dataset, we first inferred the fingerprint or embedding vector. Then the entire HARIBOSS ligand set including the known binders was ranked according to the distances between the small molecules' representative vectors and the inferred vector. For each ligand, We defined the ranking score to be the percentage of the ligands ranking below, that is to say 1 meaning the ligand was ranked as the best and 0 meaning the worst. We thus judged the model performance by the ranking score distribution of all the true binders in each case.

For the case specific screening task, we set up a pipeline to virtually screen the R-bind library starting from 3D structures using our model. We extract pockets from the structures with RLDock[21] and inferred the favorable fingerprint by RNAmigos or our E3NN model, or the Uni-Mol representation by the E3NN model. Then we screen Rbind database by distance to obtain the aforementioned ranking score. We used the HCV IRES domain II (PDB ID 2NOK) [22] and HIV-1 TAR (PDB ID 1UUI) [23] as test cases, of which 11 and 14 binders have been reported [24, 25, 26] in the R-Bind database, respectively.

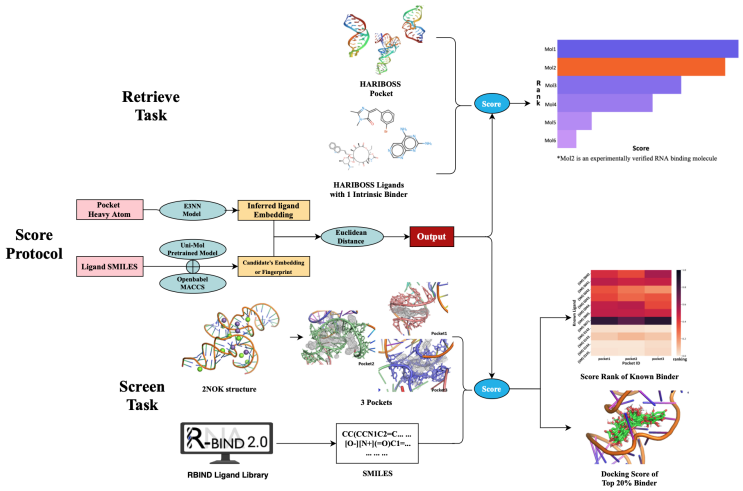

Figure 3: RNA based retrieve and screening task

3 Results

3.1 RNA-binding Ligand Discriminator

Due to the specific physical properties and the dynamic nature of RNA, the RNA binding ligands share unique yet complicated chemical similarities. Such similarities may be beyond the capabilities of classical descriptors but more suitable for sophisticated representations like Uni-Mol[12]. Indeed with the Uni-Mol embedding, our classification model can distinguish RNA binding ligands with high accuracy (Figure 4). On the augmented set of ROBIN, our model achieved an AUC value of 0.985 in the ROC curve and 0.883 in the PRC curve, both outperformed the MLP model in the original work.[5]

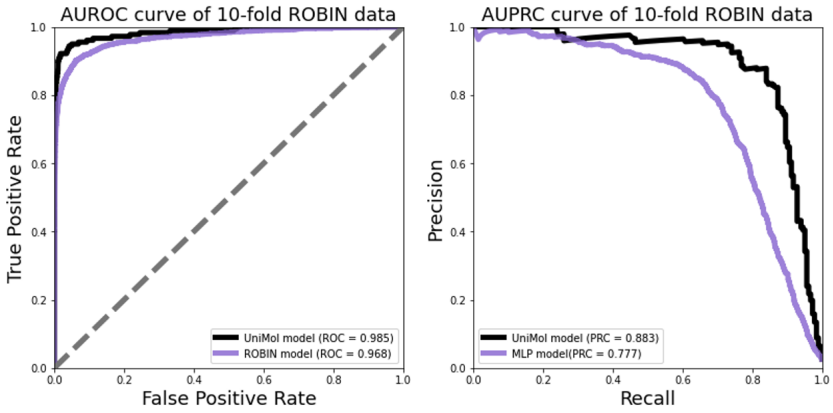

Figure 4: ROC curve and PRC curve in ROBIN's augmented data set

3.2 Uni-Mol RNA-binding QSAR

To further demonstrate the power of Uni-Mol embedding in RNA related tasks. We applied the Uni-Mol QSAR using a case study conducted by Cai et. al. [13], and compared with the original fingerprint-based QSAR model of 435 descriptors. With the same multiple linear regression head, Uni-Mol representation achieved higher correlation owing to its powerful representative ability (Table 1, Figure 5). Such a setup leaves room for more in-depth applications like QSAR-based molecule generation in the Uni-Mol framework to effectively guide lead optimization.

Table 1: Comparison of Model Performance

	pk_{on}		pk_{off}	
Data	train	test	train	test
MSE	2.3	1.46	0.51	0.35
R^2	0.77	0.85	0.83	0.93
R^2 in *Cai et. al.*	0.77	0.77	0.64	0.61

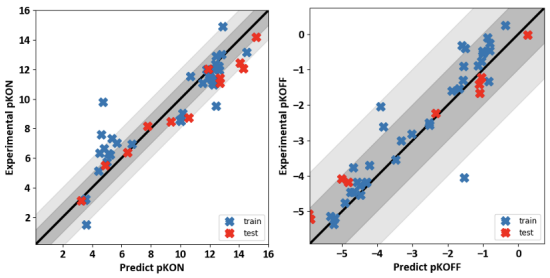

Figure 5: QSAR Regression curve

3.3 Pocket-based Molecular Representation Inference

RNAmigos[10] proposed an augmented base pairing network to infer ligands' MACCS fingerprint. In contrast, the E3-Equivariant network makes use of all atoms in the vicinity of the ligand instead of reduced presentation. Such essential and versatile description allows components other than standard nucleic acid, e.g. methyls or ions, which are often important for ligand binding, to be considered. When using the same MACCS fingerprint, our model achieved an MSE of 0.0049,0.048 and 0.041 in the train, valid, and test set respectively, compared to an MSE loss of 0.15 in RNAmigos. In ligand retrieve task. A mean ranking of 0.70 was achieved on the test set, slightly higher than RNAmigos(0.68), although notably, we used a slightly larger dataset. When converting to Uni-Mol representation, the mean ranking improved dramatically to 0.927 (Figure 6). Indicating the effectiveness of the Uni-Mol representation.

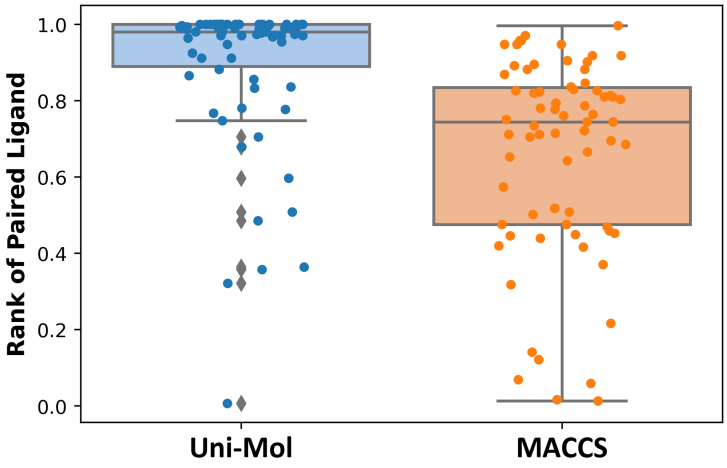

Figure 6: Box and scatter plot of retrieve task

We choose two well-studied systems to test our model for real screening tasks. One is the HCV IRES subdomain IIa with 11 known ligands included in the R-BIND dataset. Another is HIV-1 TAR RNA with 14 reported ligands in the R-BIND database.

We utilized RLDock [21] to identify potential binding sites, and three putative binding sites were identified for each system (Figure 7). Inference was then conducted on all these pockets with RNAmigos model+ MACCS representation, E3NN model + MACCS representation, and E3NN model + Uni-Mol representation.

In the HCV case, the ranking of the known binder inferred by the E3NN model is significantly higher than that inferred by RNAmigos, with an AUC in the recall-threshold curve of 0.49,0.33, respectively. The improvement is more significant under the Uni-Mol representation, with a mean ranking of 0.63 and an AUC of 0.68. Taking 20% as the criteria to screen hits, MACCS representation could only find 1 ligand in both model, while Uni-Mol identified 5 ligands. Trends in the HIV case are similar, with the AUC of 0.50,0.45 and 0.75, and the hit ratio of 1/14,3/14,7/14 for RNAmigos model+MACCS, E3NN model+MACCS, and E3NN model+Uni-Mol, respectively. We further calculated the binding affinity of top 20% hit molecules with AutoDock Vina[27]. In the HIV case,the Uni-Mol and MACCS shows comparable docking score distribution (average -6.87 vs -6.89) but for the HCV case, the docking score of Uni-Mol's hits is significantly better than that of MACCS score(-8.66 vs -7.66). It is worth noting an NMR structure without co-factors was used for the HIV case, and more canonical binding pockets were captured in the groove, making it harder for traditional docking to distinguish the true binders. But the E3NN+Uni-Mol can nevertheless successfully identify half of the true binders. These results suggest the combination of E3NN model and Uni-Mol representation can be a promising tool in future RNA targeting virtual screening.

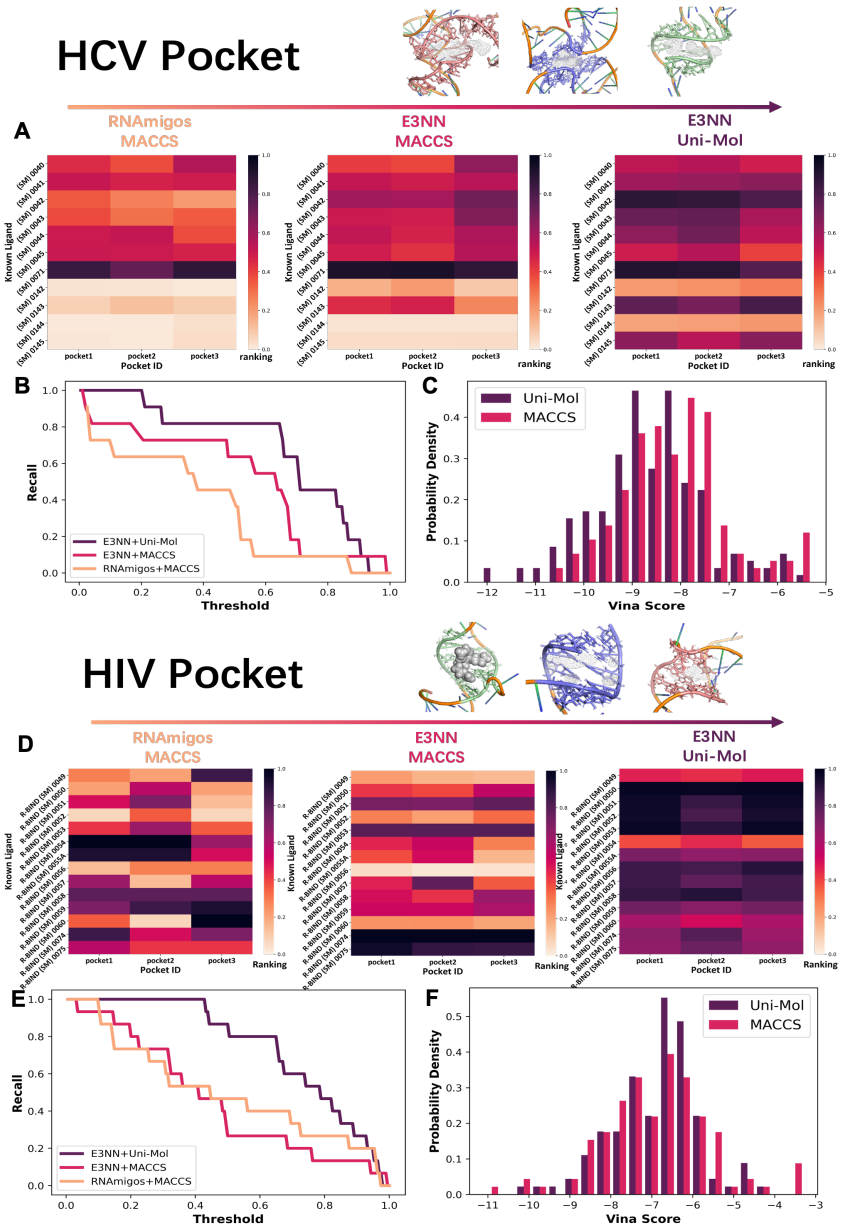

Figure 7: **A/D** are the ranking heatmaps for 11 in HCV and 14 in HIV known ligands vs. all pockets for RNAmigos MACCS, E3NN MACCS and E3NN Uni-mol model respectively; **B/E** shows the Recall-Threshold curves of different models; **C/F** shows the Vina score distributions for Uni-Mol and MACCS screening ligands; after model screening. The putative pockets for each systems shown on the top.

4 Conclusion

There is an emerging trend of RNA-targeting drug design in recent years. But limited data and difficulty in establishing its biological assays greatly hindered the RNA-targeting small molecule drug development. Here we present a deep learning framework towards rational RNA-targeting drug design in an E3-equivariant context. The ligand based model can be used to virtually screen the RNA targeting libraries and for efficient RNA orientated QSAR model. The usage of Uni-Mol embedding naturally comes with multiple benefits: within the same representation space, most of the Uni-Mol derivative downstream models, like ADMET prediction, molecule generation, can be directly used to

further broaden the computational capacities. In addition, the pocket based model can predict optimal small molecule representation given the pocket geometry, which can be directly used to construct small size libraries in combination with fingerprint or Uni-Mol similarity search or generative model. Traditional techniques like Docking and MD simulations can also be integrated into the pipeline. With even more data available, we believe our novel approach can greatly accelerate the RNA drug designs and help unlock a spectrum of new drug targets.

References

[1] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews Drug discovery*, 1(9):727–730, 2002.

[2] Qingliang Li, Tiejun Cheng, Yanli Wang, and Stephen H Bryant. Pubchem as a public resource for drug discovery. *Drug discovery today*, 15(23-24):1052–1057, 2010.

[3] Chi V Dang, E Premkumar Reddy, Kevan M Shokat, and Laura Soucek. Drugging the'undruggable'cancer targets. *Nature Reviews Cancer*, 17(8):502–508, 2017.

[4] Anita Donlic, Emily G Swanson, Liang-Yuan Chiu, Sarah L Wicks, Aline Umuhire Juru, Zhengguo Cai, Kamillah Kassam, Chris Laudeman, Bilva G Sanaba, Andrew Sugarman, et al. R-bind 2.0: An updated database of bioactive rna-targeting small molecules and associated rna secondary structures. *ACS Chemical Biology*, 2022.

[5] Kamyar Yazdani, Deondre Jordan, Mo Yang, Christopher R Fullenkamp, Timothy EH Allen, Rabia T Khan, and John S Schneekloth. Machine learning informs rna-binding chemical space. *bioRxiv*, 2022.

[6] Asher Mullard. Fda approves rna-targeting small molecule. *Nature Reviews Drug Discovery*, 19(10):659–660, 2020.

[7] Saisai Sun, Jianyi Yang, and Zhaolei Zhang. Rnaligands: a database and web server for rna–ligand interactions. *Rna*, 28(2):115–122, 2022.

[8] Francesco Paolo Panei, Rachel Torchet, Herve Menager, Paraskevi Gkeka, and Massimiliano Bonomi. Hariboss: a curated database of rna-small molecules structures to aid rational drug design. *bioRxiv*, 2022.

[9] Kaili Wang, Yiren Jian, Huiwen Wang, Chen Zeng, and Yunjie Zhao. Rbind: computational network method to predict rna binding sites. *Bioinformatics*, 34(18):3131–3136, 2018.

[10] Carlos Oliver, Vincent Mallet, Roman Sarrazin Gendron, Vladimir Reinharz, William L Hamilton, Nicolas Moitessier, and Jérôme Waldispühl. Augmented base pairing networks encode rna-small molecule binding preferences. *Nucleic acids research*, 48(14):7690–7699, 2020.

[11] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.

[12] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2022.

[13] Zhengguo Cai, Martina Zafferani, Olanrewaju M Akande, and Amanda E Hargrove. Quantitative structure–activity relationship (qsar) study predicts small-molecule binding to rna structure. *Journal of medicinal chemistry*, 65(10):7262–7277, 2022.

[14] Noel M O'Boyle, Chris Morley, and Geoffrey R Hutchison. Pybel: a python wrapper for the openbabel cheminformatics toolkit. *Chemistry Central Journal*, 2(1):1–7, 2008.

[15] Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12):2562–2574, 2015.

[16] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

[17] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

[18] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018.

[19] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022.

[20] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), may 2022.

[21] Li-Zhen Sun, Yangwei Jiang, Yuanzhe Zhou, and Shi-Jie Chen. Rldock: a new method for predicting rna–ligand interactions. *Journal of chemical theory and computation*, 16(11):7173–7183, 2020.

[22] Sergey M Dibrov, Hillary Johnston-Cox, Yi-Hsin Weng, and Thomas Hermann. Functional architecture of hcv ires domain ii stabilized by divalent metal ions in the crystal and in solution. *Angewandte Chemie*, 119(1-2):230–233, 2007.

[23] Ben Davis, Mohammad Afshar, Gabriele Varani, Alastair IH Murchie, Jonathan Karn, Georg Lentzen, Martin Drysdale, Justin Bower, Andrew J Potter, Ian D Starkey, et al. Rational design of inhibitors of hiv-1 tar rna through the stabilisation of electrostatic "hot spots". *Journal of molecular biology*, 336(2):343–356, 2004.

[24] Punit P Seth, Alycia Miyaji, Elizabeth A Jefferson, Kristin A Sannes-Lowery, Stephen A Osgood, Stephanie S Propp, Ray Ranken, Christian Massire, Rangarajan Sampath, David J Ecker, et al. Sar by ms: discovery of a new class of rna-binding small molecules for the hepatitis c virus: internal ribosome entry site iia subdomain. *Journal of medicinal chemistry*, 48(23):7099–7102, 2005.

[25] Maia Carnevali, Jerod Parsons, David L Wyles, and Thomas Hermann. A modular approach to synthetic rna binders of the hepatitis c virus internal ribosome entry site. *ChemBioChem*, 11(10):1364–1367, 2010.

[26] Jeet Chakraborty, Ajay Kanungo, Tridib Mahata, Krishna Kumar, Geetika Sharma, Ritesh Pal, Khondakar Sayef Ahammed, Dipendu Patra, Bhim Majhi, Saikat Chakrabarti, et al. Quinoxaline derivatives disrupt the base stacking of hepatitis c virus-internal ribosome entry site rna: reduce translation and replication. *Chemical Communications*, 55(93):14027–14030, 2019.

[27] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.