# Leveraging 2.5D Graph Representations and Relational Graph Convolutional Networks for RNA-Ligand Binding Site Prediction


## Abstract

Ribonucleic acids (RNAs) are widely studied as drug targets for therapies in the field of medicine. However, the intensity and cost of identifying their functions experimentally have brought on efforts to solve this problem computationally. Machine learning tools to predict the binding site of small molecules with RNAs remain a challenge. In this paper, we present a novel approach to predict small-molecule binding sites by training deep graph learning models on augmented graph representations of RNAs. To do so, we first assemble a dataset of small-molecule binding RNAs whose 3D structures are known and transform them into augmented graph representations. Further, we find that these graphs, augmented with non-canonical base pairing information (NCBPI) and 3D modules uncover a signal for binding site location. Finally, we identify and discuss some particular issues pertaining to this prediction task.


## Introduction

A study by Warner, Hajdin, and Weeks (2018), identifies approximately 70% of the human genome as non-coding RNAs (ncRNAs) which play potentially important roles in gene expression tasks (Mattick and Makunin 2006). RNAs were thought to be principally involved in transcription, but ncRNAs can take on a variety of important roles in cellular function, which makes them an attractive drug target for medicinal therapies. Adding on to that interest is the observation that small molecule-RNA binding can regulate or even halt ncRNA activity. Unfortunately, there are significant labor and material costs and challenges to determining RNA structure, function, and small binding molecule identity. Following the success of bioinformatics tools in predicting protein-ligand binding interactions, a wave of tools has emerged to fill the gap that exists in predicting RNA-ligand binding interactions. Unlike proteins, however, RNA-ligand interactions cannot be determined based on protein surface affinity, and require more complex techniques.

Working towards the overarching goal of predicting RNA function and RNA-ligand binding interactions, this large problem is broken down into subproblems. The focus of this paper is the particular subproblem of ligand binding site identification in RNA structure.



## Background

### RNA Structure

RNAs have complex structures that are organized hierarchically. The base level, called sequence or primary structure, is formed of covalent bonds between nucleotides which constitute the backbone of the molecule. Then, the secondary structure is shaped by canonical base pairing interactions between nucleotides which lead to folding patterns known as loops and stacks. While stacks are stable structures, loops can house NCBPIs between any pair of nucleotides. These NCBPIs and interactions between loops give rise to the tertiary structure of RNAs. They are composed of 3D motifs which are stereotyped across RNAs. Finally, several RNA monomers can come together to form a multimer.

An abundant amount of information can be predicted algorithmically from the primary structure. Namely, with experimentally determined binding energies of canonical base pairing interactions, secondary structure elements can be predicted using algorithms such as RNAfold (Lorenz et al. 2011). One limitation of this algorithm is that its predictions become increasingly unreliable for sequences longer than 200 base pairs in length. Nevertheless, predictions of secondary structure elements (SSEs) from RNAfold inform predictions about more complex 3D structures in RNAs. Indeed, RNA-MoIP (Reinharz, Major, and Waldispühl 2012), BayesPairing (Sarrazin-Gendron et al. 2019) and its successor BayesPairing2 (Sarrazin-Gendron et al. 2020), and CaRNAval (Reinharz et al. 2018) receive SSE predictions as input to infer tertiary structure elements, such as where 3D modules can integrate into the 3D structure of the molecule.

Data on the 3D structure of RNAs, is publicly available in the Protein Data Bank (Berman et al. 2000). RNAs can be mined from this database which contains mmcif files with the experimentally established 3D coordinates of each nucleotide in the RNA. Such information can be used to determine 12 types of NCBPIs based on the relative orientations of the pairs in 3D, as put forward by Leontis and Westhof (2001). These interactions are of paramount importance as they colocalize with binding sites (David-Eden, Mankin, and Mandel-Gutfreund 2010) and convey a signal

of ligand binding preference (Oliver et al. 2020). Conveniently, the DSSR suite (Lu, Bussemaker, and Olson 2015) directly extracts this information from mmcif files for downstream use. Further, loops, which are at the heart of many 3D modules, are enriched in non-canonical base pairs and are believed to have a functional role in RNA interactions (Sarrazin-Gendron et al. 2019).

### Graph Representations of RNA

Finding a way to integrate all this information is difficult. Graphs emerge as a natural way of representing RNA molecules since they are lower dimensional than using 3d coordinates, they are more abstract, and an abundance of tools already exist to manipulate them. Hence, directed graphs where nucleotides are encoded as nodes and all types of base pair interactions are encoded as multi-relational edges work very well. Indeed, previous works have used augmented base pairing networks (ABPNs) (Oliver et al. 2020; Reinharz et al. 2018; Sarrazin-Gendron et al. 2019; 2020; Su, Peng, and Yang 2021; Wang and Zhao 2020) to represent RNAs, but all for different prediction tasks, thereby attesting to the flexibility of ABPNs.

Central to our research, the *rnaglib* python library (Mallet et al. 2022) streamlines the process of building, annotating, and training models on ABPNs, or 2.5D graphs, as they will be referred to throughout. *rnaglib* represents RNAs as described above, and also allows for the addition of desired node and edge level attributes to augment graphs at will. Finally, the library also integrates Relational Graph Convolution Networks (RGCNs) to extract meaningful interactions between base pairs in node-, edge-, or graph-level prediction tasks.

### Ligand Binding Site Prediction

Earlier works in the field are few in number, but steady progress is being made. RSite is a geometric approach that computes the distance curve of each node relative to each other from an input tertiary structure. Their model predicts the binding site as the extremums of the curve, but they only report results on two ncRNAs (Zeng et al. 2015). Its successor, RSite2 does roughly the same computation on secondary structure inputs and reports 60% precision on 10 binding sites (Zeng and Cui 2016). Another work, RBinds (Wang and Zhao 2020), predicts RNA-ligand binding site by utilizing a thresholding mechanism based on closeness and degree of bases in a 2.5D graph. It achieves 0.82 accuracy on a test set of 22 RNAs they assembled in a previous study (Wang et al. 2018). However, they do not disclose how they compute accuracy which is problematic, as discussed later. Finally, the current state of the art, to our knowledge, is RNAsite (Su, Peng, and Yang 2021) which feeds a structure based and a sequence based model into a random forest algorithm to output a binary classification prediction for each node in the input graph, indicating the node's participation in an RNA-ligand binding site. Interestingly, their structural model builds off the 2.5D graph operations from RBinds with the additional features of Laplacian Norm to characterize surface convexity and solvent accessibility. Their statistical analyses reveal the importance of these topological features in the success of their predictions. However, the performances they report are separated on two datasets that are both broken down into easy and hard subsets, and none of them match with the numbers used in their comparison with other models' performance.

## Contribution

As mentioned above, most current tools are tested on minuscule datasets, so it is difficult to tell if they can generalize to a broader range of ncRNAs. We attempt to remedy this through our contribution which is threefold; (1) we create a dataset of RNAs that is much larger than those previously designed for similar tasks and detail the pipeline for obtaining such structures, (2) we present a novel approach to node level binary classification of binding site identification using deep graph learning on 2.5D graphs and train a model that successfully interprets ligand binding site information, and (3) we identify and discuss some specific challenges of the RNA-ligand binding site prediction problem.

## Materials and Method

Oliver et al. (2020) found that 2.5D graphs augmented with NCBPI encode small-molecule binding preferences. Therefore, we have reason to think that NCBPIs provide a signal of where ligand binding sites are positioned. Further, recurrent RNA 3D modules were found to be frequent functional sites in RNA interactions (Sarrazin-Gendron et al. 2020). Hence, we want to encode NCBPIs, 3D module information, and the SSEs on top of which the 3D modules are formed into 2.5D graphs. Accordingly, we first build a pipeline that creates a dataset of 2.5D graphs augmented with the latter features from a set of input mmcif files. Subsequently, we use this model to train a deep graph learning model that predicts binding sites given an input graph or mmcif file. The goal would be for the output to be directly usable by a program that predicts ligand binding preference, e.g. RNAmigos (Oliver et al. 2020).

### Dataset Preparation

Data scarcity has been an issue in previous RNA-ligand binding site predicting models (Wang et al. 2018; Zeng et al. 2015; Zeng and Cui 2016; Wang and Zhao 2020). We attempt to solve this problem by constructing a much larger dataset of 2.5D graphs representing RNAs with their associated binding small molecules. The dataset creation process is broken down into the following steps:

1. **Data collection:**
   Two non-redundant RNA molecule datasets are assembled: the RNAmigos dataset (Oliver et al. 2020) and the BGSU RNA 3D structures dataset (Leontis and Zirbel 2012). All duplicate structures are removed. RNAmigos originally has around 750 structures and BGSU has 899, for a total of around 1650 RNAs and over 2000 individual chains. The files come as a combination of pre-annotated 2.5D *networkx* graphs (Hagberg, Swart, and S Chult 2008) and mmcif files retrieved directly from the PDB. A first bottleneck is discarding all the RNAs that do not have a small-molecule binding site. To stay consistent

with the current literature, small molecule binding sites are defined as the ensemble of nodes falling within a 4-5 Å distance threshold of the ligand (Su, Peng, and Yang 2021; Mallet et al. 2022). Further, the *rnaglib* ligand list is used to define what ligands are valid, this does not include water, ions, spermine, or any other ligand used for experimental determination of 3D structure.

2. **Conversion to 2.5D graphs annotated with NCBPIs and SSEs:**
   First, mmcif files are fed to DSSR (Lu, Bussemaker, and Olson 2015) and turned into directed, multi-relational edge graphs by *rnaglib*. Further, RNAfold, which is integrated to the *rnaglib* API, also predicts SSEs and directly annotates graph nodes with an attribute that includes the name of the predicted SSE (e.g. hairpin, bulge, internal loop, etc.). A second bottleneck is introduced here, whereby RNAs with one or more chains containing incomplete node-level information are stripped of said chains.
3. **Chain separation:**
   The dataset is then partitioned into sub-graphs of their original complex structures into the individual chains composing them. To stay consistent with previous literature, all chains whose length falls outside the interval [20; 1500] are discarded. As mentioned above, binding pockets are typically defined as the nodes that are within 4-5 Å of the ligand, but since the relational graph model struggles with the scarcity of positively annotated nodes under the constraints of this definition, two other sets of 2.5D graphs are created. The cutoff distances defining binding sites were set to 10 and 15 Å respectively (see results for more details). Also, some RNAs with multiple chains have some that qualify and others that are filtered out. In this case, the chains that qualify are kept. However, concerns arose that some binding sites could be composed of nucleotides originating from different chains, of which only certain conform to the above criteria, which would leave the model to learn node and edge features for an incomplete binding site, and potentially affect performance. To address this issue, a separate set of monomeric RNAs is set aside, called the easy dataset.
4. **3D module prediction and node-level module annotation:**
   Of all remaining chains, those with sequences containing non-nucleotide identities are discarded. Sequences are then fed to BayesPairing2 (Sarrazin-Gendron et al. 2020) which returns the positions of predicted 3D modules and their associated scores. This score combines the probability of the sequence folding into a structure that accommodates that module with the probability of observing that module based on the nucleotide code. The identity of the module and its score are integrated to the above 2.5D graphs as node level attributes of the appropriate nucleotides. Chains with no output module predictions or with no output predictions scoring above the hyperparameter t=4 are discarded.

The final number of datasets produced is five; the set of graphs with distance threshold d = 5Å (5cutoff), the set of graphs with d = 10Å (10cutoff), the set of graphs with d = 15Å (15cutoff), the set of graphs with d= 15Å that come from a monomer (*easy*), and the set of graphs in 15cutoff that aren't in *easy* (hard). Table 1 details the information for each dataset.

| Dataset Name | Cardinality | Positive/Negative Node Ratio |
|---|---|---|
| 5cutoff | 312 | 0.0087 |
| 10cutoff | 351 | 0.0885 |
| 15cutoff | 452 | 0.1204 |
| *easy* | 46 | 0.1085 |
| *hard* | 406 | 0.1192 |

Table 1: Dataset information.

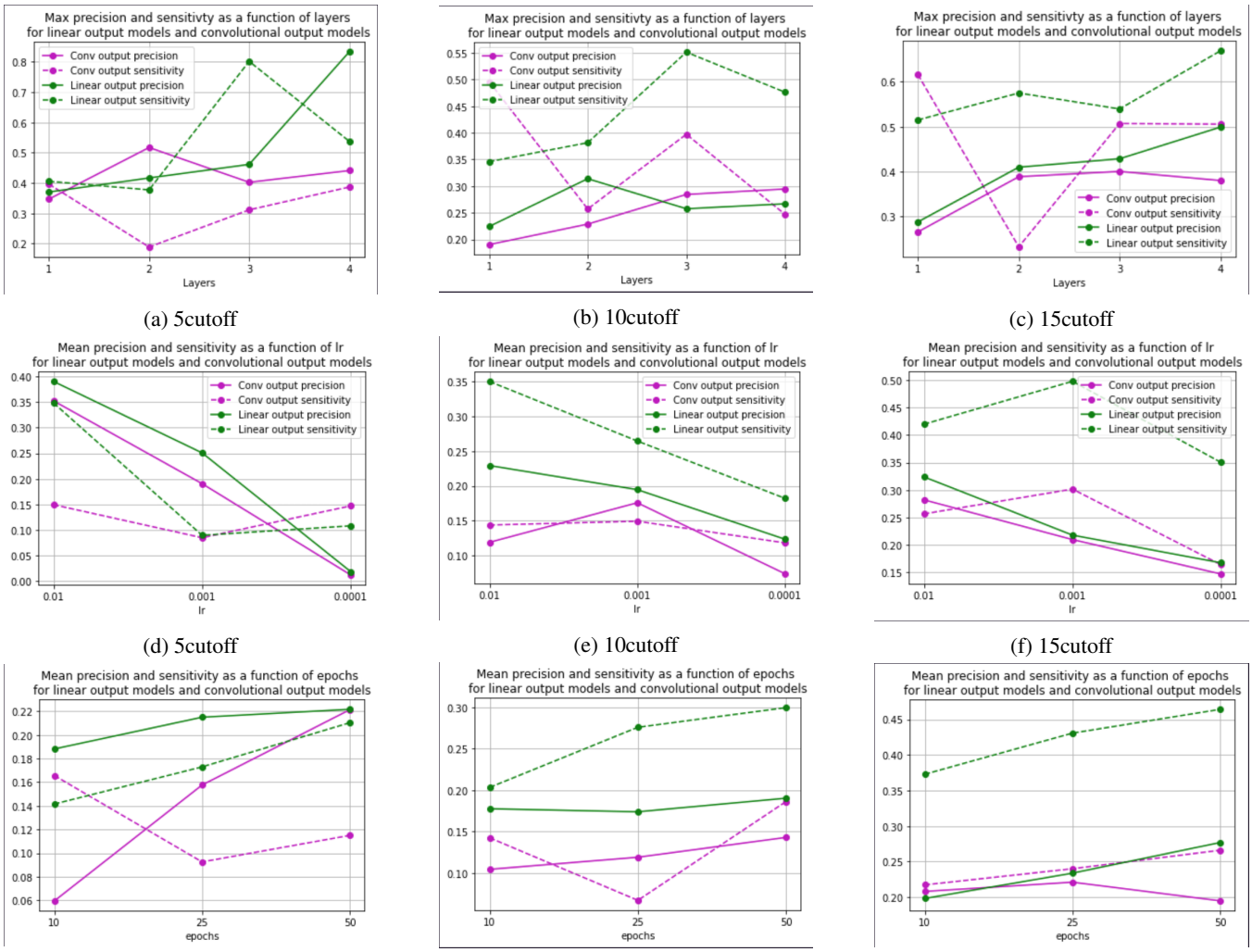
For unknown reasons, some RNAs cannot successfully be converted into 2.5D graphs and others cannot be annotated such that the output 3D module predictions match the right nodes, but this concers only a relatively small number of chains, so it is ignored for now. After all this processing, the initial 2000+ chains are reduced to 312, 351, and 452 chains for 5cutoff, 10cutoff, and 15cutoff, respectively. The difference in sizes of these datasets is mostly accounted for by the observation that many RNAs in the original 2000 do not have any nodes that qualify under the constraining 5 Å definition. However, the looser 10 and 15 Å binding site definition allows more chains to be considered.

In table 1, the difference in the difficulty of positive node prediction between the 5cutoff graph set and the 10cutoff and 15cutoff sets is apparent. Namely, the two latter sets have a greater number of graphs to be trained on. However, the main difference is the percentage of nodes that are positive samples. The 10cutoff set has 10 times as many positive samples as the 5 cutoff, and the 15cutoff set has almost 14 times more.

## Deep Graph Network

The choice of working with an RGCN model is motivated by the work of Oliver et al. (2020). The authors choose RGCNs because they extract information from multi-relational edges such as those in 2.5D graphs, as they were designed to operate on the multi-relational edges of knowledge graphs (Schlichtkrull et al. 2018). Further, the convolutional operator on graphs can pick up on 3D localization information encoded in the nodes of 2.5D graphs through SSE and 3D module attributes. Moreover, the ease of building said models through the *rnaglib* library serves as even greater reason to experiment with different RGCN architectures and parameters.

RGCNs have the property of looking deeper and deeper into a node's neighborhood as more layers are stacked on top of one another. As such, an important hyperparameter to tune is the number of layers of the model. Experiments on models with 1, 2, 3, and 4 layers are discussed later, and all models have the possibility of making the output layer linear (i.e. the 1 layer model becomes a simple 1 layer perceptron network). The learning rate and the number of training epochs are also optimized during tuning experiments.




(a) 5cutoff (b) 10cutoff (c) 15cutoff

(d) 5cutoff (e) 10cutoff (f) 15cutoff

(g) 5cutoff (h) 10cutoff (i) 15cutoff

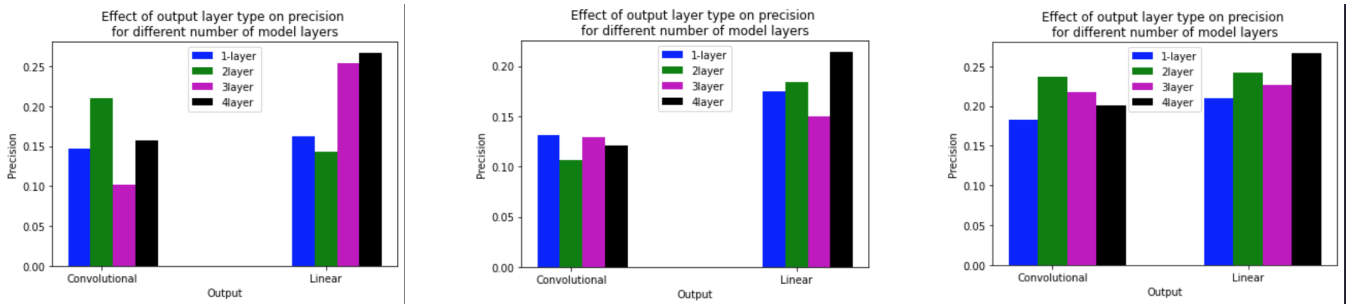
Figure 1: Average test accuracies during hyperparameter tuning on 5cutoff, 10cutoff, and 15cutoff.

We formulate the given problem of classifying each input graph's nodes as belonging to a binding site or not into a supervised training task. At testing time, model performance is reported similarly to (Su, Peng, and Yang 2021), where the totality of nodes that are correctly and incorrectly predicted are taken into account. This is different from Rsite and RBinds which report their results on a per chain basis, saying whether the binding site is recognized or not. The metrics used, for future comparison with other models, are precision and sensitivity, which are particularly oriented towards the quality of positive sample identification. The choice to not report accuracy stems from the fact that in this setting, accuracy is misleading. particularly on the 5cutoff dataset, there are as many as 99.13% of nodes in the test set that are negative samples. As such, one could create a predictor that always predicts a negative output and still obtain 99.13% accuracy. Thus, accuracy is not, and should not be used for evaluation.

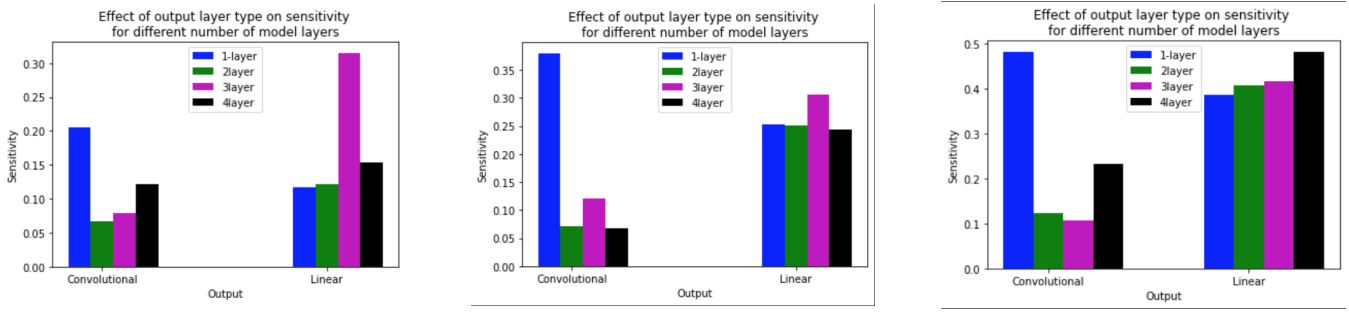
## Results

We carried out tuning experiments in a grid search fashion, performing 3 trials for each hyperparameter combination and reporting the average precision and sensitivity. Before realizing the particular imbalance in the ratio of positive to negative nodes in the different datasets, the models were trained on the 5cutoff dataset to stay consistent with other works. The model found some success especially with the 3- and 4-layer networks with linear output layer (Figure 1a). Further, in comparison with a control model that was trained only on the nucleotide code for the 5cutoff dataset, the 3-layer model had significantly higher sensitivity (0.8, Figure 1a,2a) than the control model (0.02, data not shown). The best performing model on the 5cutoff dataset has 2 convolutional layers and an output linear layer, trained for 50 epochs with a learning rate 0.01 and achieves 0.41 precision and 0.53 sensitivity.

For the models evaluated on the 10cutoff and 15cutoff datasets, we observe a similar trend of the 3 and 4 layer mod-




(a) 5cutoff (b) 10cutoff (c) 15cutoff




(d) 5cutoff (e) 10cutoff (f) 15cutoff

Figure 2: Quantifying the effects of linear vs convolutional output layers during tuning.

els with linear output layers performing very well, especially with respect to sensitivity. The two best performing models on the 15cutoff dataset both had 4 layers and were trained for 50 epochs. However, the best one had a learning rate of 0.01 and achieved 0.49 precision and 0.53 sensitivity, while the second had a learning rate of 0.001 and achieved 0.38 precision and 0.58 sensitivity. Similarly, for the 10cutoff dataset, the best model had 3 layers and was trained for 25 epochs with 0.01 learning rate, and it achieved 0.2 and 0.53 precision and sensitivity, while the second best model achieved 0.25 and 0.47, which had 4-layers (data not shown).

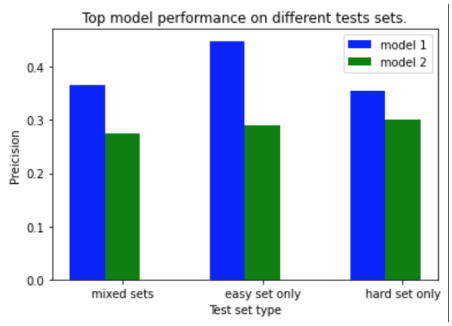
In Figures 1d-e, it is clear that learning rate has a distinct effect on precision and sensitivity. Decreasing learning rate comes with an initial boost in sensitivity at the expense of precision, and further decreasing learning rate becomes detrimental to both performance measures. These results are corroborated by the fact that none of the top performing models had a learning rate of 0.001. Moreover, it is interesting to note that while learning rate has a distinct impact on performance, the hyperparameter at the source of the greatest variation in model performance is the number of layers in the model.

We take a look at the possible effects of the output layer on the way the different models are learning. A few trends that are worth mentioning is that single layer RGCNs consistently exhibit high sensitivity, regardless of dataset (Figures 2d-f). This suggests that RGCNs are fundamentally well adapted for the type of relational data used as input in this task. This result is further corroborated by the fact that the 2nd highest performing models out of the models that have a convolutional output layer (for both the 10cutoff and 15cutoff datasets) were single layer models. However, despite this, we see that a combination of several convolutional layers with a final linear output layer yield the best results (Figure 2, last column to the right in each graph).

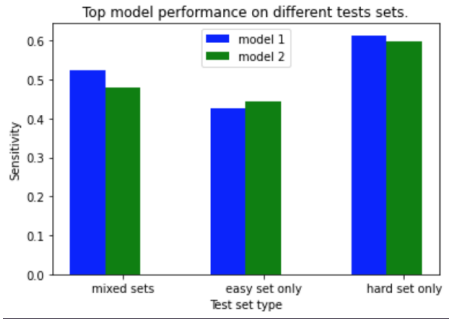
As we observed that our results improved significantly on the 15cutoff dataset, we do not do any further testing on the 5cutoff and 10cutoff datasets. Without knowledge as to why, we notice, however, that after model selection, the RGCNs demonstrate more robust generalization on the 5cutoff dataset than the 10cutoff one.

Above, we briefly mention that an additional dataset, the *easy* dataset is created. There are two motivations behind the creation of this set of graphs; (1) in case there is a significant difference in the prediction of binding sites in multimeric RNAs and monomeric RNAs, this dataset would allow us to quantify that difference, and (2), because we hypothesize that these monomers, being generally shorter would be easier to predict ligand binding sites on than larger molecules. This hypothesis stems from the observation that the quality of predictions made by RNAfold and BayesPairing2 degrades rapidly as sequence length grows past 200 base pairs. Further, in line with experiments done by (Su, Peng, and Yang 2021) (2021), the authors also evaluate their model on an easy and a hard dataset.

In another set of experiments we evaluate the performance of the top two performing models on the 15cutoff dataset and on the easy and hard datasets. In Figure 3 we summarize our results comparing the precision and sensitivity of the latter and former datasets with the original 15cutoff. The change in dataset has no effect on model 2's precision while it increases that of model 1 (Figure 3a). A more general trend




(a) Precision for 15cutoff on the easy and hard datasets




(b) Sensitivity for 15cutoff on the easy and hard datasets

Figure 3: Performance of the top two performing models on the *easy* dataset, on the *hard* dataset, and on the original 15cutoff dataset. The mixed dataset is just another name for the 15cutoff dataset. Each bar represents the average value over 5 trials.

seems to be that the easy dataset decreases model sensitivity, thus increasing the number of false negatives (Figure 3b). We thought this might suggest that the easy dataset had fewer positive samples than the hard dataset but their ratios of positive/negative nodes are too similar for that to be true (Table 1).

## Discussion and Future Work

In this paper, we tackle the challenge of predicting RNA-ligand binding sites from sequence and 3D information. Unlike any other work in this domain up until now, we create a dataset of RNA-small-molecule pairs that is scaled up 5 or 6 times compared to its competition. While this dataset is an improvement on previous works, its creation is subject to many bottlenecks that limit the final number of usable RNA chains for model training. As such, a more thorough analysis of how to rescue chains from avoidable bottlenecks so that they can fit in the pipeline would improve the dataset. Though the dataset assembled in this paper is probably our most important contribution, we also point out in the Background section how problematic reporting model performance in this field is. Namely, accuracy cannot be used as a reliable measure of performance, and a consensus on how prediction is quantified needs to be reached in this community.

We also present a novel model for binding site prediction and are the first, to our knowledge, to use RGCNs for this task. The applicability of this model is demonstrated in experiments on our dataset, showing clear signs of learning binding site locations precisely. Changing the definition of ligand binding sites by broadening the spherical radius encompassing nodes that are part of binding sites boosts model performance (only for 15Å). This means that instead of trying to identify patterns of binding site location based on 0.8% of total nodes, the model can now learn patterns in 12% of nodes. In doing so, we address the problem of having the size of positive node samples potentially hinder learning. Nevertheless, our model is still able to learn binding site locations well using the 5cutoff dataset. Our approach also strengthens the case that deep graph learning has a place in RNA structural predictions, along with RNAmigos (Oliver et al. 2020). This learning is made possible due to the compact and flexible representations of RNAs as 2.5D graphs. To boost model performance further, possible future work includes augmentation of the current graph sets with features similar to those used in Su, Peng, and Yang (2021) characterizing chain surface convexity and solvent accessibility.

Finally, for a more complete understanding of where our model ranks in terms of the competition, future work should address performance compairsons with existing models on our dataset as well as their own. A missing element of this study to improve in future work is the more thorough statistical analysis of results to corroborate our discussion.

## References

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research* 28(1):235–242.

David-Eden, H.; Mankin, A. S.; and Mandel-Gutfreund, Y. 2010. Structural signatures of antibiotic binding sites on the ribosome. *Nucleic acids research* 38(18):5982–5994.

Hagberg, A.; Swart, P.; and S Chult, D. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Leontis, N. B., and Westhof, E. 2001. Geometric nomenclature and classification of rna base pairs. *Rna* 7(4):499–512.

Leontis, N. B., and Zirbel, C. L. 2012. Nonredundant 3d structure datasets for rna knowledge extraction and benchmarking. In *RNA 3D structure analysis and prediction*. Springer. 281–298.

Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; and Hofacker, I. L. 2011. Viennarna package 2.0. *Algorithms for molecular biology* 6(1):1–14.

Lu, X.-J.; Bussemaker, H. J.; and Olson, W. K. 2015. Dssr: an integrated software tool for dissecting the spatial structure of rna. *Nucleic acids research* 43(21):e142–e142.

Mallet, V.; Oliver, C.; Broadbent, J.; Hamilton, W. L.; and

Waldispühl, J. 2022. Rnaglib: a python package for rna 2.5 d graphs. *Bioinformatics* 38(5):1458–1459.

Mattick, J. S., and Makunin, I. V. 2006. Non-coding rna. *Human molecular genetics* 15(suppl_1):R17–R29.

Oliver, C.; Mallet, V.; Gendron, R. S.; Reinharz, V.; Hamilton, W. L.; Moitessier, N.; and Waldispühl, J. 2020. Augmented base pairing networks encode rna-small molecule binding preferences. *Nucleic acids research* 48(14):7690–7699.

Reinharz, V.; Soulé, A.; Westhof, E.; Waldispühl, J.; and Denise, A. 2018. Mining for recurrent long-range interactions in rna structures reveals embedded hierarchies in network families. *Nucleic acids research* 46(8):3841–3851.

Reinharz, V.; Major, F.; and Waldispühl, J. 2012. Towards 3d structure prediction of large rna molecules: an integer programming framework to insert local 3d motifs in rna secondary structure. *Bioinformatics* 28(12):i207–i214.

Sarrazin-Gendron, R.; Reinharz, V.; Oliver, C. G.; Moitessier, N.; and Waldispühl, J. 2019. Automated, customizable and efficient identification of 3d base pair modules with bayespairing. *Nucleic acids research* 47(7):3321–3332.

Sarrazin-Gendron, R.; Yao, H.-T.; Reinharz, V.; Oliver, C. G.; Ponty, Y.; and Waldispühl, J. 2020. Stochastic sampling of structural contexts improves the scalability and accuracy of rna 3d module identification. In *International Conference on Research in Computational Molecular Biology*, 186–201. Springer.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer.

Su, H.; Peng, Z.; and Yang, J. 2021. Recognition of small molecule–rna binding sites using rna sequence and structure. *Bioinformatics* 37(1):36–42.

Wang, H., and Zhao, Y. 2020. Rbinds: a user-friendly server for rna binding site prediction. *Computational and Structural Biotechnology Journal* 18:3762–3765.

Wang, K.; Jian, Y.; Wang, H.; Zeng, C.; and Zhao, Y. 2018. Rbind: computational network method to predict rna binding sites. *Bioinformatics* 34(18):3131–3136.

Warner, K. D.; Hajdin, C. E.; and Weeks, K. M. 2018. Principles for targeting rna with drug-like small molecules. *Nature reviews Drug discovery* 17(8):547–558.

Zeng, P., and Cui, Q. 2016. Rsite2: an efficient computational method to predict the functional sites of noncoding rnas. *Scientific reports* 6(1):1–9.

Zeng, P.; Li, J.; Ma, W.; and Cui, Q. 2015. Rsite: a computational method to identify the functional sites of noncoding rnas. *Scientific Reports* 5(1):1–5.