AUGMENTED BASE PAIRING NETWORKS ENCODE RNA-SMALL MOLECULE BINDING PREFERENCES.

A PREPRINT

Carlos Oliver
School of Computer Science, McGill University,
Mila - Quebec Artificial Intelligence Institute
carlos.gonzalezoliver@mail.mcgill.ca

Vincent Mallet
Institut Pasteur, Structural Bioinformatics Unit,
MINES ParisTech, PSL Research University, CBIO - Centre for Computational Biology,
v.mallet96@gmail.com

Roman Sarrazin Gendron
School of Computer Science, McGill University,
roman.sarrazingendron@mail.mcgill.ca

Vladimir Reinharz
Department of Computer Science, Université du Québec à Montréal,
vreinharz@gmail.com

William L. Hamilton
School of Computer Science, McGill Universety,
Mila - Quebec Artificial Intelligence Institute
wlh@cs.mcgill.ca

Nicolas Moitessier
Department of Chemistry, McGill University
nicolas.moitessier@mcgill.ca

Jérôme Waldispühl*
School of Computer Science, McGill University,
jeromew@cs.mcgill.ca

January 31, 2020

ABSTRACT

Motivation: The binding of small molecules to RNAs is an important mechanism which can to stabilize 3D structures or activate key molecular functions. To date, computational and experimental efforts toward small molecule binding prediction have primarily focused on protein targets. Considering that a very large portion of the genome is transcribed into non-coding RNAs but only few regions are translated into proteins, successful annotations of RNA elements targeted by small-molecule would likely uncover a vast repertoire of biological pathways and possibly lead to new therapeutic avenues.
Results: Our work is a first attempt at bringing machine learning approaches to the problem of RNA drug discovery. RNAmigos takes advantage of the unique structural properties of RNA to predict small molecule ligands for unseen binding sites. A key feature of our model is an efficient representation of binding sites as augmented base pairing networks (ABPNs) aimed at encoding important structural

*Corresponding author.

patterns. We subject our ligand predictions to two virtual screen settings and show that we are able to rank the known ligand on average in the 73rd percentile, showing a significant improvement over several baselines. Furthermore, we observe that graphs which are augmented with non-Watson Crick (a.k.a non-canonical) base pairs are the only representation which is able to retrieve a significant signal, suggesting that non-canonical interactions are an necessary source of binding specificity in RNAs. We also find that an auxiliary graph representation task significantly boosts performance by providing efficient structural embeddings to the low data setting of ligand prediction. `RNAmigos` shows that RNA binding data contains structural patterns with potential for drug discovery, and provides methodological insights which can be applied to other structure-function learning tasks.
Availability: Code and data is freely available at `http://csb.cs.mcgill.ca/RNAmigos`.
Contact: `jerome@cs.mcgill.ca`

1 Introduction

Recent studies have identified small organic molecules as important non-covalent regulators of RNA function in many cellular pathways [1]. These discoveries contribute to a better understanding of key pathways present in all organisms, but also pose RNA molecules as a large class of promising novel drug targets. For example, Ribocil, which has recently been uncovered through a phenotypic assay to target the FMN riboswitch, is currently undergoing clinical trials as a novel antibiotic [2]. Various other small molecule-activated RNA systems are also being proposed [3, 4, 5]. Notable among these is the application to CRISPR activation regulation [6]. The list of possible therapies is likely to expand given the observations of KD Warner and co-workers that only a small fraction of the genome is translated into protein (1.5%) while the vast majority is transcribed into potentially druggable non-coding RNA (70%) [7].

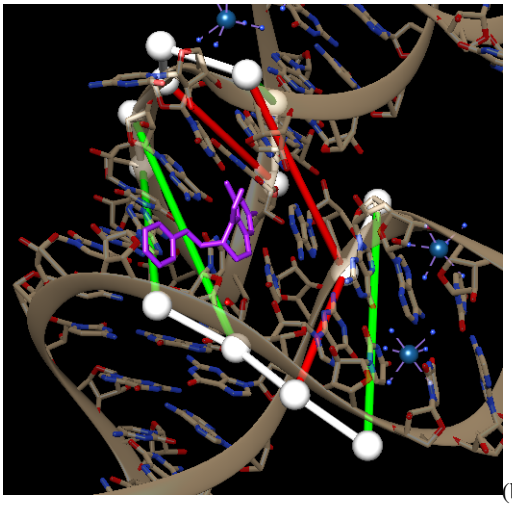
In parallel, the protein-binding drug discovery field is experiencing rapid and significant advances from data-driven artificial intelligence models. Tasks such as candidate molecule generation and affinity scoring [8] which were previously done using costly knowledge-based simulations are now highly automated [9]. One of the largest points of resistance for data-driven RNA drug discovery is the relatively small number of available binding assays and crystallized RNA-small molecule complexes (34,271 solved protein structures with a ligand, vs 2,253 for RNA). While the number of solved RNA structures is steadily growing, successful models for automated RNA 3D function annotation will likely require customized methods which carefully leverage domain knowledge. To this end, we turn to current knowledge of RNA structural organization to motivate base pairing networks as a strong prior in low-data settings.

1.1 RNA Structural Organization

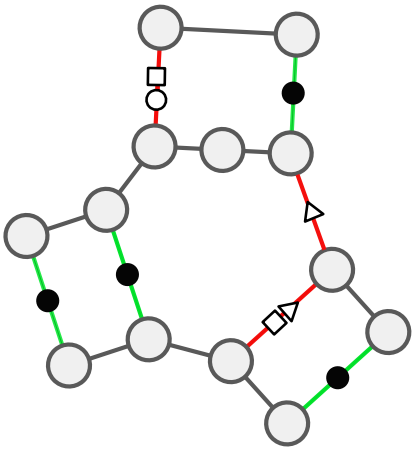
RNAs possess multiple levels of structural organization which together determine function, and by extension, ligand binding specificity. At the simplest level, RNA is a string of monomers {`A,U,C,G`} linked by a chain of covalent bonds known as the backbone. This is commonly known as as the primary structure of RNA. Non-covalent pairwise interactions between nucleotides (bases) in the chain give rise to the secondary and tertiary structure. Canonical pairs (i.e. `A-U`, `C-G`), are the most studied class of base pair, give rise to the secondary structure. Notably, canonical pairs series of nested loops and stable stacks, assembling a scaffold for the full structure [10]. The experimental determination of binding energies for these pairs [11] prompted a boom of explicit algorithms for sequence to secondary structure prediction such as `RNAfold`, [12]. In seminal work, Leontis and Westhof identified 11 additional types of base pairing occurring in 3D structures [13, 14], known as non-canonical base pairs. These interactions can occur between any pair of nucleotides and are defined by the relative orientations of three faces of the interacting bases. By considering all combinations of faces and a *cis* and *trans* orientation, we arrive at 12 possible base pairing geometries. Whereas canonical pairs form stable helices, non-canonical pairs are enriched in loops (i.e. regions without canonical pairs) and create more complex structural patterns [15, 16]. These pairings fine-tune RNA function by defining structure at the 3D level [17]. Interestingly, non-canonical pairs were also found to be enriched in ligand binding sites [18, 19], which corroborates with the observation that increased structural complexity is associated with binding specificity [7].

These observations led us to hypothesize that studying RNA structures at the augmented base-pairing level (i.e. including non-canonical pairs) holds useful spatial and chemical information about ligand binding. However, studying RNA at this level of structure comes with major algorithmic challenges, namely, binding energies of non-canonical interactions has not yet been determined, and non-canonical interactions are known to be highly non-nested (i.e. many crossing interactions). For these reasons, non-canonical interactions are typically modeled with statistical methods, and represented using more general data structures such as graphs [20] instead of trees or grammars as is the case for 2D structures. In practice, this means that a graph using vertices to represent nucleotides and multi-relational edges to encode base-pairing interactions could offer a signature for RNA ligand binding sites (See **Fig. 1** for an example of a binding site and its associated base pairing network). We call this graphical representation of RNA sites annotated with

canonical and non-canonical interactions an Augmented Base Pairing Network (ABPN) since we consider base pairs beyond the canonicals. Indeed, similar representations of RNA base pairing networks have been exploited in various tools [21, 22, 16, 20] for their ability to capture RNA-specific interactions in an interpretable manner. This paradigm distinguishes RNA from protein-ligand interactions where surface-cavity topologies tend to drive binding preferences [23], making direct use of atomic coordinates more appropriate.



(a) Binding site atomic coordinates



(b) Graph encoding of binding site as an augmented base pairing network (ABPN).

Figure 1: RNA structure representation of the THF riboswitch binding site (PDB: `4LVV`) as atomic coordinates using UCSF Chimera [24](left) and resulting augmented base pairing network (ABPN) (right). We superpose the ABPN in the 3D visualization. Nodes are drawn as white spheres, backbone connections are in white, and canonical and non-canonical base pairs are green and red tubes respectively. We color the edges simply to guide the eye to the corresponding base pairs but note that edge color has no special meaning to our graphs. We annotate the graph representation with the standard Leontis-Westhof nomenclature for pairing type symbols. In this case, the binding site has three canonical interactions denoted (●), and three non canonicals of types (□○, ▷, □▷).

1.2 Structure-based Drug Discovery and RNA Base Pairing Networks

The central aim of structure-based drug discovery is to identify compounds with high affinity to a given site or set of binding sites. A natural problem to address in this context is the prediction of binding affinity from a binding site-ligand pair. Machine learning models which solve this task can be used as alternatives to computationally expensive docking simulations to screen of ligand databases for strong binders [25]. This setting is quite feasible in the protein domain as affinities and drug screens are abundant, hence various methods have been proposed [26]. Recently, some repositories of RNA small molecule data have been made public [27] however, only a handful of binding affinities are known. Two automated affinity scoring approaches have been proposed; DrugScoreRNA, and LigandRNA [28, 29], however these rely on chemical knowledge for constructing scoring functions. While more accurate yet slower RNA docking methods are still showing limited success [23, 30]. To our knowledge, no fully data-driven methods for RNA structural drug discovery have been developed.

We therefore turn to the set of RNA-small molecule binding events captured by crystallography and made publicly available in the RCSB Protein Databank. By extracting the atomic coordinates around all organic compounds bound to RNA, we are able to obtain a set of binding site-ligand pairs for machine learning. Our model, `RNAmigos`, seeks to learn the relationship between RNA structures (represented as augmented base pairing interaction networks) and ligands (encoded as vectors of chemical features). Instead of producing a scalar affinity score for a ligand-RNA complex, we output a chemical description (fingerprint) of a ligand given a binding site. An advantage of this framing is that our model would only require as input a binding site and not an enumeration of structure-ligand pairs, as is the case in the usual affinity prediction models. Such instances of the drug discovery problem have recently been proposed in the protein domain [31, 32]. By directly yielding an active compound for each binding site, identifying ligands for large sets of sites is greatly accelerated. Given the pervasiveness of RNA transcripts, such models are likely to become especially relevant in the RNA domain.

1.3 Contribution

RNAmigos brings together domain knowledge of RNA structure, currently available crystal structure data, and graph neural networks, to show that base pairing networks can be used to automatically recover ligands for RNA structures. We propose the use of Augmented Base Pairing (ABPNs) networks which encode an enriched alphabet of base pairing interactions and demonstrate that it is a necessary component for capturing binding signatures. Molecular fingerprints predicted by RNAmigos serve as effective ligand search tools across diverse ligand classes and shows strong performance in two different ligand screens. Additionally, we explore the use of an unsupervised graph representation learning scheme for boosting model performance in this low-data setting. This approach holds promise for applications in other structure-function prediction tasks for RNA. Our model takes the first steps toward data-driven RNA drug discovery as well as establishes the groundwork for structure-function prediction of complex RNA structures. The core implementation of RNAmigos is built in Pytorch [33] and DGL [34] and is available as an open source Python 3.6 software package.

2 Methods

At a high level, RNAmigos accepts a base pairing network as input and predicts a ligand fingerprint. More formally, we learn a mapping from node and edge-attributed graphs $G = (V, E, \mathcal{V}, \mathcal{E})$ to binary vectors $\hat{y} \in [0, 1]^k$ in a multi-label classification setting. The set of node attributes includes the 4 nucleotide types $\mathcal{V} = \{A, U, C, G\}$ and edge attributes contain the set of 12 base pairing geometries and the backbone pair, $\mathcal{E} = \{\texttt{backbone}, \{cis, trans\} \times \{\circ, \square, \triangleright\}^2\}$. This process is supervised by minimizing the difference between the predicted fingerprint and the observed fingerprint (in this case, co-crystallized ligand). Finally, we evaluate the quality of the predictions in a ligand screen by using $\hat{y}$ to search for y in a larger set of ligands (decoys). This allows us to better contextualize the raw model performance in a more realistic setting. We provide a full overview of RNAmigos in **Fig. 2**.

Here we outline the steps of training set construction, model architecture, and ligand screen validation.

2.1 Dataset Preparation

Since no benchmark datasets exist for this task, we begin by building a set of RNA-small molecule complexes from the RCSB PDB Data Bank [35]. We download all crystal structures (90% identity threshold) which contain RNA and at least one ligand. This results in 2993 PDB structures. We omit ions such as magnesium (Mg+) from the set of valid ligands as they vastly outnumber organic ligands and likely require customized models [36] We choose a maximum allowable distance between any ligand atom and any RNA atom of 10 Angstroms according to *David-Eden et al.* [18] which statistically characterized ribosome antibiotic binding sites. The number of valid sites is further reduced when we remove binding sites with fewer than five RNA residues and remove binding sites containing a large proportion of protein residues as these are unlikely to form a useful structural signal. The final training set consists of 773 binding sites associated to 270 unique ligands.

Finally, we build an ABPN from the 3D structure of each binding site identified in the previous step, where each node corresponds to a residue in the binding site, and edges are formed between edges if they form a backbone or base pair interaction. Node and edge annotations are taken from the BGSU RNA 3D Motif Atlas [16] database which maintains base pairing annotations of all PDBs with Leontis-Westhof and backbone interaction types computed by the software FR3D [37]. In this manner, each ABPN stores the nucleotide identity (A, U, C, G) of each of its residues a a node attribute, and each base pairing interaction corresponds to an edge with one of 13 different types (backbone + 12 base pairing geometries). Since we applied a hard distance cutoff in the crystal structure, the resulting ABPNs often have backbone discontinuities. To address this issue, we expand the ABPN to include nodes for residues found at 1-hop from the initial node set. Graphs are on average 15.76 nodes in size. At this point, the ligand is removed from the structure so that the graph contains only RNA base-pairing information. While atomic coordinates are the current source of data, we highlight that a key feature of taking ABPNs as input is that we can eventually learn from many other sources of data which are easier to obtain. A promising example comes from recent developments in predicting base pairing networks [22] from RNA sequences in high-throughput. Our model would then be able to directly take advantage of such data once it is linked to a functional label (such as a fingerprint in this case). For full details on binding site extraction and graph construction, see Supplementary Material.

2.2 Fingerprint Prediction

The output side of our model is a vector $\hat{y} \in [0, 1]^k$ which is trained to approximate the chemical fingerprint of the bound ligand. Many functions to compute fingerprints from chemical structures have been developed, all with the

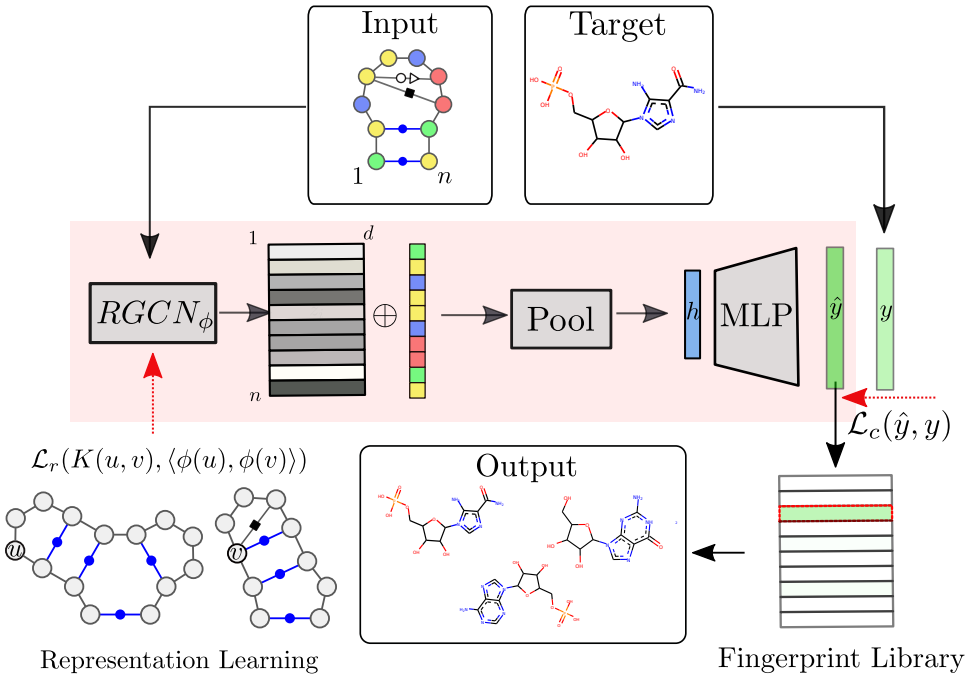


Figure 2: Outline of the `RNAmigos` pipeline. An base pairing network is passed as input to `RNAmigos`. In training mode, it is paired with a known ligand (Target) from which a target fingerprint y is constructed. The embedding network (RGCN) produces a matrix of node embeddings of dimension $n \times d$ where n is the number of nodes in the graph. This is followed by a pooling step which reduces node embeddings to a single graph-level vector. Finally, the graph representation is fed through a multi-layer perceptron (MLP) to produce a predicted fingerprint $\hat{y}$. The fingerprint is then used to search for similar ligands to the prediction in a ligand screen and thus enriches the probability of identifying an active compound. The RGCN network is pre-trained using an unsupervised node embedding framework which allows us to leverage structural patterns from a large dataset of RNA structures.

common aim of encoding chemicals in a vector space where similar compounds will lie in proximal regions of the space [38, 39, 40]. In this work we use a very common fingerprint implementation known as the MACCS fingerprint [41] which has the advantage of being compact (k = 166 vs the usual 1024) and has interpretable dimensions. For a given chemical compound c, the MACCS fingerprint f_c is a 166 bit binary vector where each dimension indicates the presence or absence of a chemical property. For the ith chemical property, $f_c[i]$ is set to 1 if the chemical property is present and is 0 otherwise. We use the set of 166 predefined chemical properties from the [42] implementation as a target vector for our model. We emphasize that the computation of the fingerprint depends only on the chemical composition of the ligand and not on the RNA binding site. In a cheminformatics setting, fingerprints are usually compared using the Tanimoto similarity, or the equivalent Jaccard distance which are simple intersections over union metrics [43] for binary vectors. However, in our machine learning setting, each fingerprint dimension is treated as multi-label category for which we output a real-valued probability. For this reason, we perform all our fingerprint comparisons using the Euclidean distance.

2.3 Model Architecture

Since a key feature of our ABPNs is the fact that we encode base paring geometry as an edge category (or relation type), we use a Relational Graph Convolutional Network (RGCN) [44] as the core of the fingerprint prediction model. An RGCN is a function $\phi_\theta : G \rightarrow \mathbb{R}^{|V| \times d}$ with learnable parameters θ which maps nodes in a graph to real vectors of size d known as node embeddings. Embeddings for each node are obtained by repeatedly applying a trainable aggregation function over the neighbourhood of each node. The relational aspect of these networks comes in the ability to account for multi-relational graphs, (graphs with edges belonging to different categories) by allowing for parameter sharing between edge types. For this work, we consider the base pairing types to be distinct categories. We believe this is a fair approximation given the results of isostericity comparisons showing that computing the geometric discrepancy between of all pairs of edge types yields close to a diagonal matrix [45].

Once node embeddings are computed, we concatenate the resulting embedding matrix with a one-hot encoding of the input graph's nucleotides (`A, U, C, G`). Next, graph-level representation is obtained by applying a widely used trainable Graph Attention Pooling layer [46], $GAT_\pi : (V, E) \rightarrow \mathbb{R}^d$ to map the node embeddings to a single vector $h \in \mathbb{R}^d$. Finally, we feed h through a Multi Layer Perceptron which yields probabilities for each index of the fingerprint y. We supervise this process using the binary cross entropy between the predicted fingerprint and the observed over all dimensions i,

$$\mathcal{L}_{fp} = \sum_{i=0}^{k} [y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{1}$$

2.4 Unsupervised Pre-Training: ABPN Node Embeddings

Since a relatively small number of RNA-small molecule binding events have been captured by crystallography, we propose an auxilary task where data is much more abundant. More specifically, we take inspiration from recent unsupervised graph representation learning techniques [47] to leverage information from the much larger set of all crystallized RNA structures (57,6060 training points vs 773 binding sites). Instead of supervising our model using only feedback from the fingerprints, we can build an RGCN with parameters π' which has learned to efficiently represent local RNA structure as vector embeddings. If this training is successful, the fingerprint prediction network can use the unsupervised RGCN weights θ' as a richer initialization for the RGCN in the supervised setting (See **Fig. 2** for an illustration of this process). In the unsupervised setting, we simply require that the embeddings for a pair of nodes $\phi_\theta(u)$ and $\phi_\theta(v)$ respect a user-defined similarity relation.

We are free to choose the pairwise similarity function $K : (u, v) \rightarrow [0, 1]$ according to domain knowledge,.

We adapt the node similarity function proposed in `struc2vec` [48] which allows us to capture local structural similarity across graphs. Other node kernels such as the ones used in `GraphClust` for RNA 2D structures [49] are only able to compare nodes within the same graph and are affected by the distance between nodes which is not necessarily related to structure. Our similarity function addresses these limitations by comparing the distribution of edge types in the local neighbourhood of u and v. We define $K_L(u, v)$ as

$$k_L(u, v) := N^{-1} \sum_{l=0}^{L-1} \lambda^l d(R_u^l, R_v^l) \tag{2}$$

Where d is a function which compares the sets of edges edges at a distance l, from u and v, denoted R_u^l, R_v^l. We use N as a normalization constant ($= \frac{1-\lambda}{1-\lambda^L}$) to ensure the sum saturates at 1. The λ^l is a decay term which allows us to attend more to structural information close to the root nodes and we set $\lambda = 0.5$. In our case, we define d to be a simple overlap measure on the histograms of base-pairing edge types f_R, and f_R' (i.e $f_R(i)$ stores the number of times edge type i is observed).

$$d(R, R') := \frac{|f_R \cap f_R'|}{|f_R \cup f_R'|} \tag{3}$$

To compensate for the over-representation of a few edge types such as backbones and Watson-Crick edges, we scale the d value with the commonly-used Inverse-Document Frequency (IDF) factor [50].

The quality of the representations is evaluated using an L2 loss between the similarity value in the graph space and the cosine similarity of the embeddings,

$$\mathcal{L}_{rep} = ||K(u, v) - cosine(\phi_\theta(u), \phi_\theta(v))||_2^2 \tag{4}$$

The inductive nature of graph neural networks allows us to simply input the obtained weights Φ to the different task of fingerprint prediction network. In this manner, we inject general RNA structure information in the supervised learning process where data is scarce.

2.5 Ligand Screen

Finally, we place the raw performance measure of the model ($\mathcal{L}_{fp}$, similarity to the observed ligand fingerprint) in the context of a ligand screen. In a ligand screen, one is given a set of compounds $\mathcal{C}$ and a target. The goal is to rank ligands in $\mathcal{C}$ by likelihood to be active binders. Here we place our predictions in such a setting. Since no external benchmarks exist for the task of fingerprint reconstruction, this lets assess whether predictions are close enough to the true value to be useful. Given a predicted fingerprint $\hat{y} \in [0,1]^{166}$ for a binding site and a library of fingerprints $\mathcal{C} \cup y$, the screen ranks all $y' \in \mathcal{C}$ by similarity to $\hat{y}$. A good $\hat{y}$ will assign a high rank to y, the true binder and thus is presumed to be an effective prediction for $\mathcal{C}$. For comparison purposes, we compute a normalized rank as follows:

$$\text{rank}_{\mathcal{C}}(y, \hat{y}) = 1 - \frac{\rho_{y,\hat{y},\mathcal{C}}}{|\mathcal{C}|} \tag{5}$$

Where $\rho_{y,\hat{y},\mathcal{C}}$ is the un-normalized rank of the true ligand y in $\mathcal{C}$ relative to the prediction $\hat{y}$. A successful predictor will rank the true ligand as closest to its prediction (normalized rank close to 1), while a random predictor will result in an average rank of 0.5.

Considering that the distribution of RNA ligands appears to cluster to specific sub-regions (see Supp. Fig 2), this readout also ensures that a classifier does not obtain a good score by simply predicting the average ligand as it would with a raw distance between y and $\hat{y}$.

In benchmarking settings $\mathcal{C}$ is typically known as a decoy set where the set is assumed to entirely of non-binders except for a single binder, which we hope to be able to detect. We construct two decoy sets for our experiments. Since there are currently no experimentally validated data sets of active and inactive binders for a given RNA site (such as DUDE for protein [51]), our first set consists of all RNA-binding ligands (270 ligands) in the PDB as our ligand library. The second decoy set is constructed using *DecoyFinder* [52] on default settings, which samples a list of 36 decoys for each compound such that generic chemical properties are preserved while potentially disturbing binding potential. Of course, this test assumes that chemical dissimilarity between an active compound $c \in \mathcal{C}$ implies inactivity which is not always the case [53]. However, the current aim of our work is simply to determine whether ABPNs retain significant amount of information about its observed ligand preferences, for which this test is sufficient.

3 Results

We evaluate our performance using the rank metric defined in Equation 5. We report resulting rank over the list of all RNA-small molecule pairs as well as the set of all decoys for the each ligands, following the two decoy benchmark process.

Due to the limited size of our labeled data set, we performed a 10-fold cross-validation to include all training pairs in the evaluation and provide a more accurate measure of performance.

Node embeddings are computed using a 3-layer RGCN, each layer consisting of 16 dimensional inputs and outputs, a graph attention layer computing a 16 dimensional graph embedding and a fully-connected layer, which outputs a 166-dimensional vector. See Supplemental Table 1 for full model architecture and hyperparameters. Variations of the architecture used did not have strong effects on performance, so no extensive hyper-parameter search was conducted. We leave the exploration of other architecture choices for future work.

3.1 Augmented RNA Base Pairing Networks Encode Binding Preferences

Setting : The first hypothesis to test is that the proposed framework (*ABPN*) is able to retrieve some signal. To explore this question, we compute the rank and distance metrics on corrupted data. We compare this performance to three baselines:
· *random* consists of a synthetic label set where each binding site is assigned a uniformly random 166-dimensional binary vector.
· *swap* is designed to account for imbalances in the data (some ligands are more frequent than others): each binding site is assigned a fingerprint selected at random from the set of observed fingerprints. The overall distribution of ligand fingerprints thus remains the same but the input-output correlations are broken.
· *majority* is a constant ligand annotation computed as a majority vote over all fingerprints at each index. This is to be compared to the *swap* to check that the only thing that can be learnt on swapped data is over-representation of some ligands within the experiment.

Experiment	Ranks		L2	
	DecoyFinder	RNA	*DecoyFinder*	RNA
random	0.611	0.542	0.502	0.502
majority	0.621	0.603	0.175	0.179
swap	0.617	0.603	0.177	0.179
no-label	0.628	0.606	0.176	0.180
primary	0.624	0.592	0.181	0.186
secondary	0.631	0.605	0.178	0.182
ABPN	0.695	0.681	0.155	0.160
ABPN + unsup.	**0.735**	**0.715**	**0.145318**	**0.150189**

Table 1: Mean ligand screen ranks and L2 distance achieved on held-out binding sites for each experiment and decoy set.

The distributions of performance over each input-output pair is visualized for all experiments in **Fig. 3** as a kernel density estimate and summary statistics can be found in **Table 1** with accompanying standard deviations in Supp. Table 2. We also assessed the statistical significance of the difference of the means in a pairwise Wilcoxon rank test which is shown in **Table 2**.

Performance : In the RNA-ligands setting, our full model achieves a rank of 0.68 and an mean-squared error (MSE) of 0.150 to the true fingerprint. The *random*, *swap* and *majority* experiments respectively yield ranks of 0.542, 0.603 and 0.603 and MSEs of 0.5, 0.18 and 0.18. This confirms that this model retrieves signal for the data and outperforms baselines. This conclusion is statistically significant based on a Wilcoxon p-value of at most $7e^{-18}$ between the model and the randomized results. As expected, the majority scheme is statistically equivalent to the swapped one and superior to the random one. These results are similar in the *DecoyFinder* setting where the mean rank of the model is 0.69 compared to 0.62 in the majority setting. This shows that the full model successfully retrieves some signal and beats the baselines.

3.2 Augmented Base Pairing Networks Encode Ligand Binding Information

We want to test the hypothesis that robust descriptors in the form of ABPN from the RNA domain knowledge are key to retrieve this signal. The question is whether the non-canonical interactions encode information that lower levels of structure (secondary, primary) do not. We answer this question by performing 3 ablation experiments on our training set:
· *primary* encodes the binding site as graphs that only contain node sequence and backbone interactions.
· *secondary* uses only information from the secondary structure which includes canonical pairs and backbones.
· *no-label* preserves all the interactions (and thus graph structure) in the graph (including non-canonical) but do not distinguish between different edge types (i.e. edges only have one label).
In all these conditions, we find that **performance is no better than the randomized baselines**, indicating that non-canonical interactions are essential for encoding specificity in ligand binding. Indeed, the best performing model is *no-label* which has a Wilcoxon p-value of 0.55 with the *majority* experiment and of $1.7e^{-18}$ with the ABPN. This finding is in agreement with biological literature on RNA binding sites and the importance of complex structural motifs for determining functional specificity [18, 7]. This is a major validation of the hypothesis that these are the correct representation for RNA structure for this task.

3.3 Unsupervised Pre-Training Boosts Performance

As explained before, one major limitation for this supervised task is the paucity of data. We investigated the possibility of using unsupervised learning by leveraging pre-trained weights on an unsupervised task as described above, and denote this experiment as *ABPN unsup*. The use of unsupervised pre-training of the node embedding network provides a significant performance boost over a network trained only on fingerprint reconstruction (MSE = 0.68 vs MSE=0.715), with a p-value of $2.9e^{-6}$ This is a methodological insight that can have applications for various other RNA-related tasks for which labeled data is typically scarce.

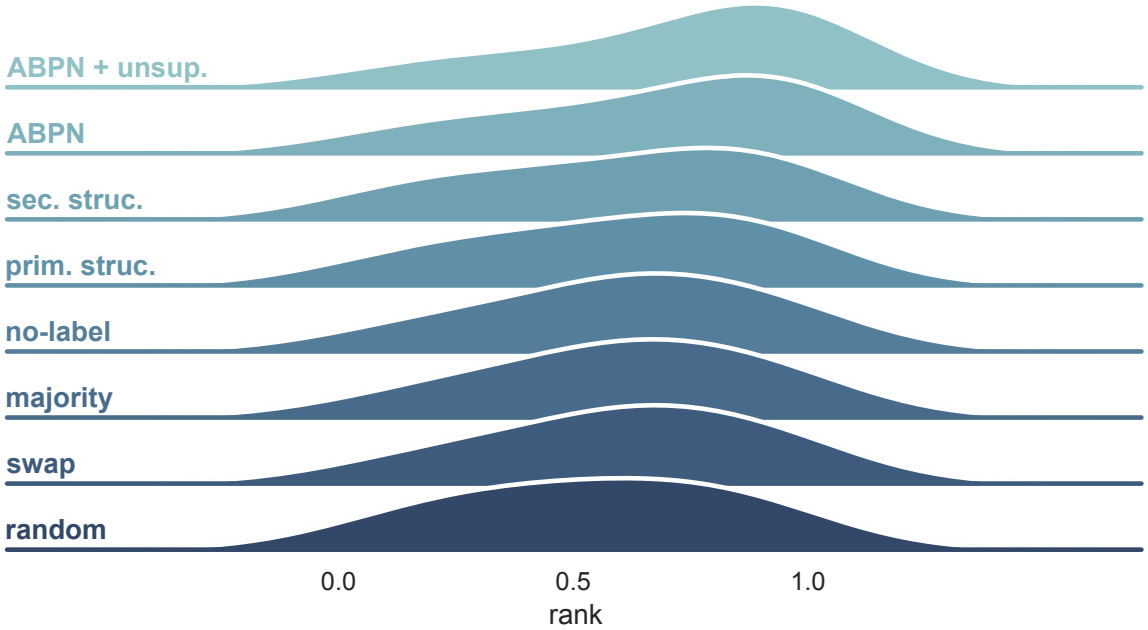


Figure 3: Distribution of L2 distances to the true ligand and rank achieved on ligand screening. Densities are estimated using KDE and represent the distribution achieved on all held-out binding sites during 10-fold cross-validation.

experiment 2 experiment 1	aBPN	secondary	primary	no-label	majority	swap	random
aBPN + unsup.	**2.9e-06**	**5.1e-26**	1.4e-22	2.1e-21	9.3e-25	7.2e-26	2.3e-18
aBPN	-	1.7e-11	5.6e-11	1.5e-08	4.3e-10	6.4e-12	2.0e-08
secondary		-	3.2e-01	7.7e-01	1.3e-01	2.8e-02	1.7e-01
primary			-	4.3e-01	2.7e-01	2.4e-02	3.2e-01
no-label				-	5.5e-01	1.5e-02	1.8e-01
majority					-	3.7e-01	3.3e-01
swap						-	5.5e-01

Table 2: Wilcoxon rank test for all pairs of training conditions. Each entry in the table is the p-value for testing the hypothesis that the ranks resulting from a pair of experiments come from the same distribution. These are performed on the RNA decoy set. We provide the test results for the *DecoyFinder* decoy set in Supplemental Table 3 material and show consistent results.

3.4 Our Model Can Predict Diverse Ligand Classes

Next, we ask whether the positive results can be explained by a small set of ligands, or whether it is able to achieve high scores on a diverse set of ligands. To get a better view of performance, we plot the same prediction scores but averaged over ligand types (270 unique ligands) against a hierarchical clustering dendrogram of each ligand (shown in **Fig. 4**).

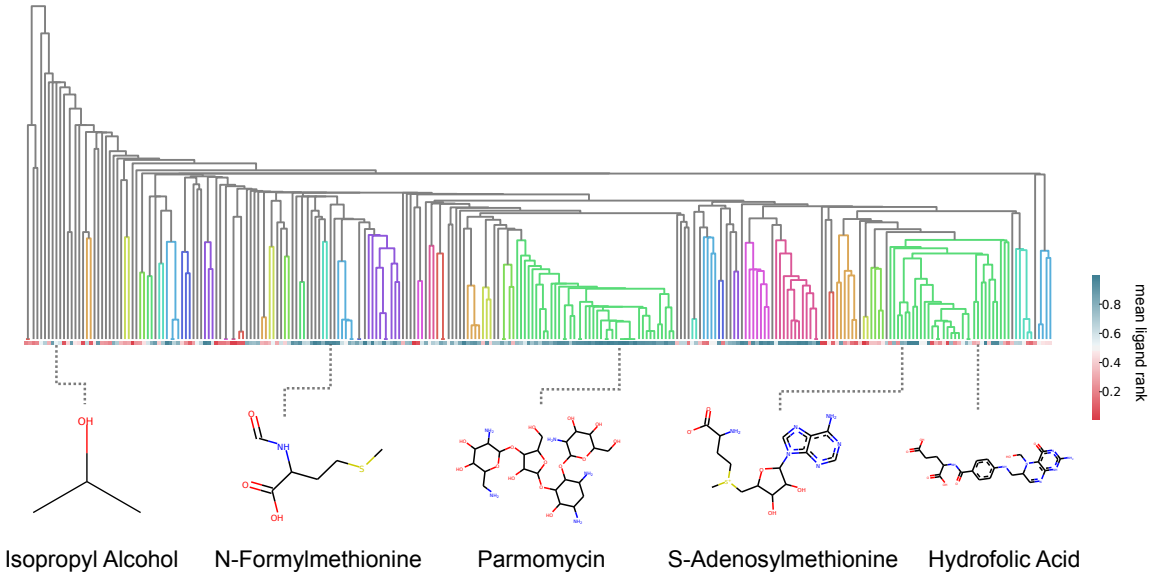


Figure 4: RNAmigos performance by ligand class. Hierarchical clustering dendrogram of the ligands, classifying ligand families by similarity. Each cell in the horizontal grid is the average score for binding sites containing a given ligand. Ligands belonging to the same tree are grouped together by the clustering procedure. Colored-in sub-trees denote tight clusters which contain ligands within 0.25 Jaccard distance.

Colored-in subtrees indicate groups of ligands that are similar, (i.e., within 0.25 Jaccard distance of each other) which would indicate strong clustering. In this manner, we are able to assess the performance across 'classes' of similar ligands. We first observe that successful classifications are not restricted to a single class of ligands and instead show good predictions for diverse ligands. Interestingly, the class that is most consistently predicted accurately corresponds to the aminoglycosides (highlighted in the pink cluster to the right). Aminoglycosides are a class of antibiotics binding to bacterial RNA with well-defined binding sites[54], and are quite abundant in the dataset. Nucleic acid-like compounds, many of which bind riboswitches, also form a large family of binders (green) however results were less consistent than for aminoglycosides. A possible explanation for strong performance on aminoglycosides, apart for the larger number of examples obtained, is that these are typically large polysaccharide-like structures with a large number of interactions with the RNA. On the other hand, riboswitches bind much smaller molecules with a limited number of interactions. As a result, binding site requirements are much more complex and specific with aminoglycosides and the large number of interactions can only be fulfilled by a limited number of molecules. Finally, ligands clustered on the left of the dendrogram show the weakest performance. Since these groups show little branching in the dendrogram, we can conclude that they represent sparsely populated ligand classes for which we have few examples and thus, obtaining more data in these regions could improve performance.

4 Discussion

We have developed a unique computational platform, `RNAmigos`, to show that augmented RNA base pairing networks contain useful ligand binding information. The significance of our results is two-fold.

We show for the first time that ABPNs encode sufficient information for a classification task, and establish an initial methodological primitives for such a task. To date, the majority of computational methods which leverage ABPNs have focused on sequence to structure [55, 22] prediction and motif identification [21, 16]. While these tasks involve some

degree of learning, the relevance of higher-order interactions lies ultimately in their potential to specify function, which until now has been left unexplored. Interestingly, these findings come at a time when information of the type our model uses is becoming more widely available. Computational prediction tools such as [22] promise to yield large amounts of higher-order RNA pairwise interaction data without need for costly crystallography experiments. This opens the door to applying such data in other important biological problems such as RNA binding protein prediction [56] and ion binding [36]. Furthermore, the promising results obtained from the unsupervised pre-training provide a methodological building block for assisting in supervised learning on complex RNA structures.

Second, our findings take an initial step towards data-driven methods for systematically identifying drugs binding to RNA, and pinpoint ABPNs as essential for this task. The finding that only ABPNrepresentations of binding sites was able to produce a significant signal in the task indicates that richer representations are necessary for successful classification when complex interactions are at play. Since our prediction is a fingerprint vector (chemical descriptor) and not a simple classification of ligands (i.e directly selecting a single ligand as output, or predicting an affinity), the fingerprint itself can be used as a tool for searching large ligand databases. While performance was strong across different ligand classes, it is apparent that classes for which data is more abundant received more consistently positive predictions. Therefore, as more examples of RNA-ligand complexes are characterized by experimental and computational techniques, we believe that the performance of our platform will improve. Additional data will also allow for considerations regarding properties desired in medical applications such as synthesizability and drug-likeness [57]. Our choice of graphs for binding site representation reflects this consideration, as graphs can natively hold additional information such as evolutionary or chemical without requiring changes to the pipeline. Eventually, computational predictions of ABPNs from sequence [22] combined with our methods will enable transcriptome-scale searches for binding sites.

We hope that this work will motivate further investigation of the links between ABPNs and RNA function, and eventually facilitate efforts in RNA targeted drug discover.

Acknowledgements

The authors would like to thank Mathieu Blanchette and Jacques Boitreaud for helpful feedback and discussions.

Funding

This work has been supported by the Natural Sciences and Engineering Research Council of Canada [RGPIN-2015-03786, RGPAS 477873-15] to JW, Genome Canada [BCB 2015] and Canadian Institutes of Health Research [BOP-149429] to JW and NM, and as well the Quebec Fonds de recherche Nature et technologies through graduate scholarships to CO. and RSG.

References

[1] Anita Donlic and Amanda E Hargrove. Targeting rna in mammalian systems with small molecules. *Wiley Interdisciplinary Reviews: RNA*, 9(4):e1477, 2018.

[2] John A Howe, Hao Wang, Thierry O Fischmann, Carl J Balibar, Li Xiao, Andrew M Galgoci, Juliana C Malinverni, Todd Mayhood, Artjohn Villafania, Ali Nahvi, et al. Selective small-molecule inhibition of an rna structural element. *Nature*, 526(7575):672, 2015.

[3] Tyler E Wagner, Jacob R Becraft, Katie Bodner, Brian Teague, Xin Zhang, Amanda Woo, Ely Porter, Bremy Alburquerque, Brian Dobosh, Oliwia Andries, et al. Small-molecule-based regulation of rna-delivered circuits in mammalian cells. *Nature chemical biology*, 14(11):1043, 2018.

[4] Ely B Porter, Jacob T Polaski, Makenna M Morck, and Robert T Batey. Recurrent rna motifs as scaffolds for genetically encodable small-molecule biosensors. *Nature chemical biology*, 13(3):295, 2017.

[5] Simone Rauch, Krysten A Jones, and Bryan Dickinson. Small molecule-inducible rna-targeting systems for temporal control of rna regulation in vivo. 2020.

[6] Kale Kundert, James E Lucas, Kyle E Watters, Christof Fellmann, Andrew H Ng, Benjamin M Heineike, Christina M Fitzsimmons, Benjamin L Oakes, Jiuxin Qu, Neha Prasad, et al. Controlling crispr-cas9 with ligand-activated and ligand-deactivated sgrnas. *Nature communications*, 10(1):2127, 2019.

[7] Katherine Deigan Warner, Christine E Hajdin, and Kevin M Weeks. Principles for targeting rna with drug-like small molecules. *Nature Reviews Drug Discovery*, 17(8):547, 2018.

[8] Erik Gawehn, Jan A Hiss, and Gisbert Schneider. Deep learning in drug discovery. *Molecular informatics*, 35(1):3–14, 2016.

[9] Sean Ekins, Ana C Puhl, Kimberley M Zorn, Thomas R Lane, Daniel P Russo, Jennifer J Klein, Anthony J Hickey, and Alex M Clark. Exploiting machine learning for end-to-end drug discovery and development. *Nature materials*, 18(5):435, 2019.

[10] Ignacio Tinoco Jr and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[11] Susan M Freier, Ryszard Kierzek, John A Jaeger, Naoki Sugimoto, Marvin H Caruthers, Thomas Neilson, and Douglas H Turner. Improved free-energy parameters for predictions of rna duplex stability. *Proceedings of the National Academy of Sciences*, 83(24):9373–9377, 1986.

[12] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

[13] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of rna base pairs. *Rna*, 7(4):499–512, 2001.

[14] Neocles B Leontis and Eric Westhof. Conserved geometrical base-pairing patterns in rna. *Quarterly reviews of biophysics*, 31(4):399–455, 1998.

[15] Neocles B Leontis and Eric Westhof. Analysis of rna motifs. *Current opinion in structural biology*, 13(3):300–308, 2003.

[16] Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of rna 3d motifs and the rna 3d motif atlas. *Rna*, 2013.

[17] Neocles B Leontis, Aurelie Lescoute, and Eric Westhof. The building blocks and motifs of rna architecture. *Current opinion in structural biology*, 16(3):279–287, 2006.

[18] Hilda David-Eden, Alexander S Mankin, and Yael Mandel-Gutfreund. Structural signatures of antibiotic binding sites on the ribosome. *Nucleic acids research*, 38(18):5982–5994, 2010.

[19] Efrat Kligun and Yael Mandel-Gutfreund. Conformational readout of rna by small ligands. *RNA biology*, 10(6):981–989, 2013.

[20] José Almeida Cruz and Eric Westhof. Sequence-based identification of 3d structural modules in rna with rmdetect. *Nature methods*, 8(6):513, 2011.

[21] Vladimir Reinharz, Antoine Soulé, Eric Westhof, Jérôme Waldispühl, and Alain Denise. Mining for recurrent long-range interactions in rna structures reveals embedded hierarchies in network families. *Nucleic Acids Research*, 46(8):3841–3851, 2018.

[22] Roman Sarrazin-Gendron, Vladimir Reinharz, Carlos G Oliver, Nicolas Moitessier, and Jérôme Waldispühl. Automated, customizable and efficient identification of 3d base pair modules with bayespairing. *Nucleic acids research*, 47(7):3321–3332, 2019.

[23] Jiaying Luo, Wanlei Wei, Jérôme Waldispühl, and Nicolas Moitessier. Challenges and current status of computational methods for docking small molecules to nucleic acids. *European journal of medicinal chemistry*, 168:414–425, 2019.

[24] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera?a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[25] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

[26] José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

[27] 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.

[28] Patrick Pfeffer and Holger Gohlke. Drugscorerna knowledge-based scoring function to predict rna- ligand interactions. *Journal of chemical information and modeling*, 47(5):1868–1876, 2007.

[29] Anna Philips, Kaja Milanowska, Grzegorz Łach, and Janusz M Bujnicki. Ligandrna: computational predictor of rna–ligand interactions. *RNA*, 2013.

[30] Li-Zhen Sun, Dong Zhang, and Shi-Jie Chen. Theory and modeling of rna structure and interactions with metal ions and small molecules. *Annual review of biophysics*, 46:227–246, 2017.

[31] Vincent Mallet, Carlos G Oliver, Nicolas Moitessier, and Jerome Waldispuhl. Leveraging binding-site structure for drug discovery with point-cloud methods. *arXiv preprint arXiv:1905.12033*, 2019.

[32] Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. *arXiv preprint arXiv:1809.02032*, 2018.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[34] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. Deep graph library: Towards efficient and scalable deep learning on graphs. *arXiv preprint arXiv:1909.01315*, 2019.

[35] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank nucleic acids research, 28: 235-242. *URL: www. rcsb. org Citation*, 2000.

[36] Anna Philips, Kaja Milanowska, Grzegorz Lach, Michal Boniecki, Kristian Rother, and Janusz M Bujnicki. Metalionrna: computational predictor of metal-binding sites in rna structures. *Bioinformatics*, 28(2):198–205, 2012.

[37] Michael Sarver, Craig L Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B Leontis. Fr3d: finding local and composite recurrent structural motifs in rna 3d structures. *Journal of mathematical biology*, 56(1-2):215–252, 2008.

[38] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme. *IDrugs*, 9(3):199, 2006.

[39] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[40] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[41] 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.

[42] Noel M O'Boyle, Chris Morley, and Geoffrey R Hutchison. Pybel: a python wrapper for the openbabel cheminformatics toolkit. *Chemistry Central Journal*, 2(1):5, 2008.

[43] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[44] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[45] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of rna base pairs. *Nucleic acids research*, 37(7):2294–2312, 2009.

[46] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[47] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[48] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 385–394, 2017.

[49] Steffen Heyne, Fabrizio Costa, Dominic Rose, and Rolf Backofen. Graphclust: alignment-free structural clustering of local rna secondary structures. *Bioinformatics*, 28(12):i224–i232, 2012.

[50] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.

[51] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

[52] Cereto Massagué Adrià, S Garcia-Vallvé, and G Pujadas. DecoyFinder, a tool for finding decoy molecules. *Journal of Cheminformatics*, 4(S1), May 2012.

[53] Marcus Bantscheff, Arjen Scholten, and Albert JR Heck. Revealing promiscuous drug–target interactions by chemical proteomics. *Drug discovery today*, 14(21-22):1021–1029, 2009.

[54] Frank Walter, Quentin Vicens, and Eric Westhof. Aminoglycoside–rna interactions. *Current opinion in chemical biology*, 3(6):694–704, 1999.

[55] Craig L Zirbel, James Roll, Blake A Sweeney, Anton I Petrov, Meg Pirrung, and Neocles B Leontis. Identifying novel sequence variants of rna 3d motifs. *Nucleic acids research*, 43(15):7504–7520, 2015.

[56] Michael Uhl, Florian Heyl, Rolf Backofen, et al. Graphprot2: A novel deep learning-based method for predicting binding sites of rna-binding proteins. *bioRxiv*, page 850024, 2019.

[57] Christopher A Lipinski. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004.

EXTENDED RNA BASE PAIRING NETWORKS IMPRINT SMALL MOLECULE BINDING PREFERENCES

CARLOS OLIVER, VINCENT MALLET, ROMAN SARRAZIN GENDRON, VLADIMIR REINHARZ, NICOLAS MOITESSIER, JÉRÔME WALDISPÜHL

1. Data Preparation

A crucial step in our learning pipeline is the construction of ABPNs from crystal structure data. Once all crystal structures are acquired, we consider spheres of varying radii around the ligand to define a binding site. We studied two parameter choices which affect the number and quality of the extracted binding sites: radius and protein vs. RNA content. As the radius increases, we obtain a larger number of binding sites. However, since the crystal structures often contain proteins, we increase the probability that the binding site will be dominated by protein residues. We tehrefore compute the ratio of RNA to Protein residues in the binding site. The resulting counts for binding sites are shown in **Fig. ??**. From this data we choose a minimum RNA concentration of 0.6 for our training model. If a PDB contains multiple binding events of the same ligand, we keep only one at random to reduce redundancies. At this stage, we have identified a set of atomic coordinates which correspond to binding sites. Since we apply a hard distance cutoff in the crystal structure, the resulting graphs often have chain discontinuities. To address this issue, we add all 1-neighbour breadth first nodes to the original graph, as well as remove any disconnected components with fewer than 4 nodes.

2. Unsupervised Pre-Training

We show in **Fig. ??** and example of a pair of nodes that obtained a high similarity score after embeddings were computed.

3. Model Architecture and Hyperparameters

4. Results

Hyperparameter	**Value**
RGCN Layers Dimensions	16, 16, 16
RGCN Number of Relations	13
RGCN Basis Sharing	None
RGCN Activation	ReLu
RGCN Dropout Probability	0.5
GAT Layer	Default
Fully Connected Dimensions	16 166

TABLE 1. Hyperparameter choices for learning pipeline. RGCN parameters are identical for the unsupervised pre-training and the fingerprint prediction networks.

Experiment	**Ranks**		**L2**	
	DecoyFinder	RNA	*DecoyFinder*	RNA
random	0.265880	0.276721	0.0384392	0.038299
majority	0.320012	0.269375	0.073969 0.074892	
swap	0.319836	0.269233	0.071212	0.071308
no-label	0.317259	0.272816	0.072830	0.073768
primary	0.323843	0.064917	0.181	0.066853
secondary	0.318527	0.299428	0.074738	0.076667
ABPN	0.322124	0.301635	0.091479	0.092328
ABPN + unsup.	0.303712	0.294309	0.093006	0.095090

TABLE 2. Standard deviation on ligand screen ranks and L2 distance achieved on held-out binding sites for each condition on both decoy sets.

method_2 method_1	ABPN	secondary	primary	no-label	majority	swap	random
ABPN + unsup -	2.9-06	5.0e-26	1.4-22	2.0e-21	9.3e-25	7.1-26	2.3e-18
ABPN	-	1.6e-11	5.6e-11	1.4e-08	4.2e-10	6.3e-12	2.0e-08
secondary		-	3.2e-01	7.6e-01	1.2e-01	2.8e-02	1.7e-01
primary			-	4.2e-01	2.7e-01	2.3e-02	3.1e-01
no-label				-	5.5e-01	1.5e-02	1.7e-01
majority					-	3.6e-01	3.3e-01
swap						-	5.4e-01

TABLE 3. Pairwise Wilcoxon test for the DecoyFinder decoy set over the ligand ranks.

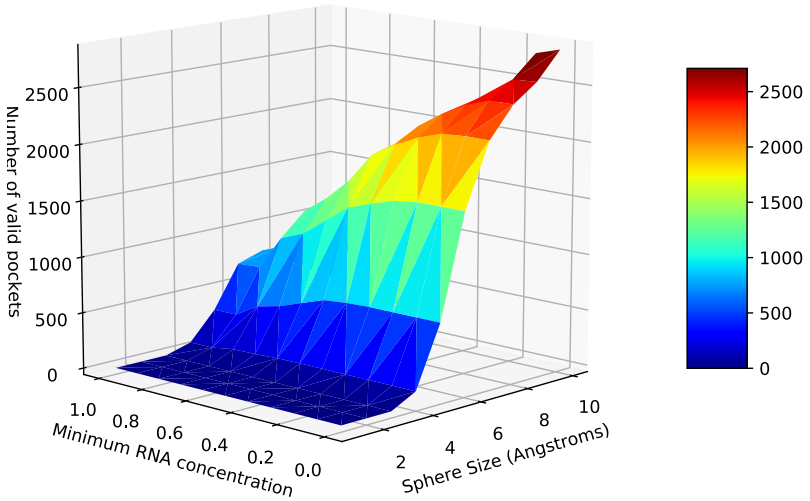


FIGURE 1. Number of binding sites retrieved versus distance threshold and RNA concentration threshold

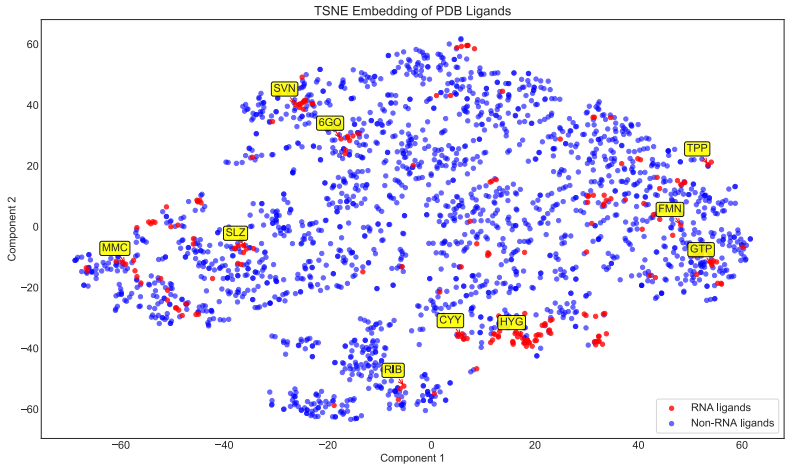


FIGURE 2. Two dimensional TSNE [**?**] embeddings of chemical fingerprints sampled from the PDB databank (RNA and protein binding). RNA ligands are highlighted in red and protein ligands in blue. We label a few interesting ligands such as 'KAN' and 'FMN' which correspond to well-known RNA binding classes known as aminoglycosides and riboswich-binding amines respectively.

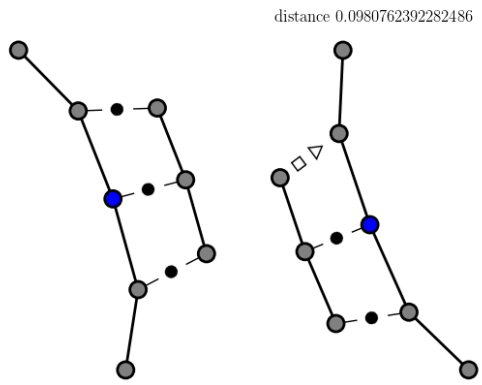


FIGURE 3. Example of pair of nodes given similar embeddings $\phi(u), \phi(v)$. The central pair of nodes which were used to make the comparison are colored in blue.