RNAmigos2: Fast and accurate structure-based RNA virtual screening with semi-supervised graph learning and large-scale docking data

Juan G. Carvajal-Patiño [†3,1], **Vincent Mallet** [†2], **David Becerra** [1,3], **Luis Fernando Niño Vasquez** [3], **Carlos Oliver** [*4,5,6], and **Jérôme Waldispühl** [*1]

[1]Department of Computer Science, McGill University, Canada
[2]LIX, Ecole Polytechnique, IP Paris, France
[3]Universidad Nacional de Colombia - Sede Bogotá, Facultad de Ingeniería – Depto. de Ingeniería de Sistemas e Industrial. Bogotá, Colombia
[4]MLSB, Max Planck Institute of Biochemistry, Germany
[5]D-BSSE, ETH Zürich, Switzerland
[6]SIB, Swiss Institute of Bioinformatics, Switzerland

[†]These authors contributed equally.
[*]To whom correspondence should be addressed.
oliver@biochem.mpg.de
jerome.waldispuhl@mcgill.ca

Abstract

RNAs constitute a vast reservoir of mostly untapped drug targets. Structure-based virtual screening (VS) methods screen large compound libraries for identifying promising candidate molecules by conditioning on binding site information. The classical approach relies on molecular docking simulations. However, this strategy does not scale well with the size the small molecule databases and the number of potential RNA targets. Machine learning emerged as a promising technology to resolve this bottleneck. Efficient data-driven VS methods have already been introduced for proteins, but these techniques have not yet been developed for RNAs due to limited dataset sizes and lack of practical use-case evaluation.

We propose a data-driven VS pipeline that deals with the unique challenges of RNA molecules through coarse grained modeling of 3D structures and heterogeneous training regimes using synthetic data augmentation and RNA-centric self supervision. We report strong prediction and generalizability of our framework, ranking active compounds among inactives in the top 1% on average on a structurally distinct drug-like test set. Our model results in a thousand-times speedup over docking techniques while obtaining higher performance. Finally, we deploy our model on a recently published *in-vitro* small molecule microarray experiment with 20,000 compounds and report enrichment factors at 1% of 8.8 to 16.8 on four unseen RNA riboswitches. This is the first experimental evidence of success for structure-based deep learning methods in RNA virtual screening. Our source code and data, as well as a Google Colab notebook for inference, are available on GitHub. [1]

[1]github.com/cgoliver/rnamigos2

Preprint. Under review.

Main

Only a small fraction of RNA codes for proteins, while ncRNA is now known to play critical roles in a wide range of biological processes [64]. For instance, about 2000 genes code for micro-RNAs, which in turn affect the expression of 60% of genes [15]. Despite this ubiquity, the first RNA-targeting drug, *risdipalm*, was only recently approved by the FDA [35], and nearly all commercially available small molecule therapies still target proteins. Being able to target RNA would drastically increase the size of the druggable space, and propose an alternative for overused protein targets in scenarios where they are insufficient. For instance, lncRNAs could represent interesting therapeutic targets in oncology [57], for which protein targets might be too specialized. RNA targets also represent a therapeutic avenue in pathologies where protein targets are absent, such as triple-negative breast cancer [66]. In this context, RNA is increasingly recognized a a promising family of targets for the development of novel small molecule therapeutics [14, 18, 11, 2], highlighting the need for efficient tools for RNA drug discovery [14].

Here, we focus on *structure-based* approaches which rely on the knowledge of the 3D structure of the RNA of interest to search for candidate compounds. A major advantage of this methodology is to allow for the discovery of binding modes beyond those already known and achieve high specificity [3]. Recently, many binding affinity prediction methods based on machine learning [4, 50, 19, 44, 61, 65, 30] were introduced, challenging traditional ones and unlocking state-of-the-art performance. Despite the widely recognized potential for RNA small-molecule therapeutics, the development of computational tools for identifying promising RNA ligands has faced substantial roadblocks compared to computational protein-targeting small molecule discovery [32]. A major reason for this lag is the substantial gap in available data of RNA structures and small molecule interactions and affinity measurements. Whereas the RCSB-PDB database [45] contains hundreds of thousands of experimental protein structures, and now hundreds of millions predicted structures from AlphaFold [22], only a couple of thousands are available for RNAs. Similarly, databases of protein-ligand interactions such as PDBBind [63] have been widely used for affinity prediction models and contain tens of thousands of experimentally measured ligand interactions for proteins and only around 100 for RNAs. In addition to the paucity of data, the uniqueness of biophysical phenomena that underpin RNA folding make it difficult to translate approaches developed for proteins into the RNA domain [56].

In recent years, several computational methods for RNA drug discovery confronted these challenges. Two major strands have emerged: docking-based and direct scoring approaches. For a given pocket-ligand pair, docking approaches estimate the binding affinity by searching over possible poses of a ligand, iteratively scoring different poses while minimizing an energy function. Widely used nucleic-acid-specific docking tools include rDock [46], Autodock VINA [58], and RLDOCK [53]. To enhance the docking procedure, alternative pose scoring functions have been proposed such as AnnapuRNA [51], LigandRNA [42], and very recently FingeRNAt [55], and RLaffinity [54]. These methods offer higher confidence screening results but come at the cost of high computational demands: the time needed to search over poses in a single binding site-ligand pair with a docking software is on the order of minutes. Therefore, instead of undergoing the expensive docking procedure, a few direct scoring methods have been proposed to estimate the binding probability without pose searching, typically using in data-driven strategies. Tools such as InfoRNA [12] and InfoRNA2.0 [13] score pocket-ligand pairs directly according to the secondary structure similarity to a fixed database of experimentally determined RNA-ligand complexes. Our previous contribution RNAmigos [37] aimed to eliminate the need of a fixed ligand library and made the first attempt at integrating the *tertiary structure* of the binding site into deep learning assisted compound screening, showing the potential of graph representations based on the Leontis-Westhof nomenclature of base pairs in 3D [27], and outperforming InfoRNA2.0 at native ligand recovery. *However, at present, no structure-based RNA virtual screening tool combines the accuracy of docking with the scalability of direct scoring methods.*

In this paper, we introduce the first structure-based method for virtual screening on RNA that is competitive with molecular docking in a fraction of the time, opening the door to large-scale target-based RNA drug discovery. Whilst our previous work provided insights on RNA binding sites representation as graphs [37], the paucity of RNA small molecule binding data impeded the performance and generalization of this approach, hindering its use in real drug discovery pipelines. We address this data paucity by generating a large database of docking scores, together with unsupervised data, as data augmentations. Moreover, we improve our RNA binding site graph representations

and propose novel model designs that open the door to enhanced neural networks and pretraining techniques. Despite running in seconds instead of hours, the proposed approach retrieves higher enrichment factors to molecular docking (top 1.1% vs 4% of the candidate ligand list). Moreover, using the tool and docking jointly, we are able to enhance docking virtual screening enrichments by cutting computation costs tenfold while improving the predicted rank of actives from 1.1% to 0.6%, This strong performance is found to be maintained over a diverse and unseen set of targets and compounds. Finally, we achieve strong enrichment factors on an independent experimentally derived ligand screen of 20,000 compounds which constitutes the first evidence of success for structure-based deep learning assisted virtual screening for RNA.

Results

Overview

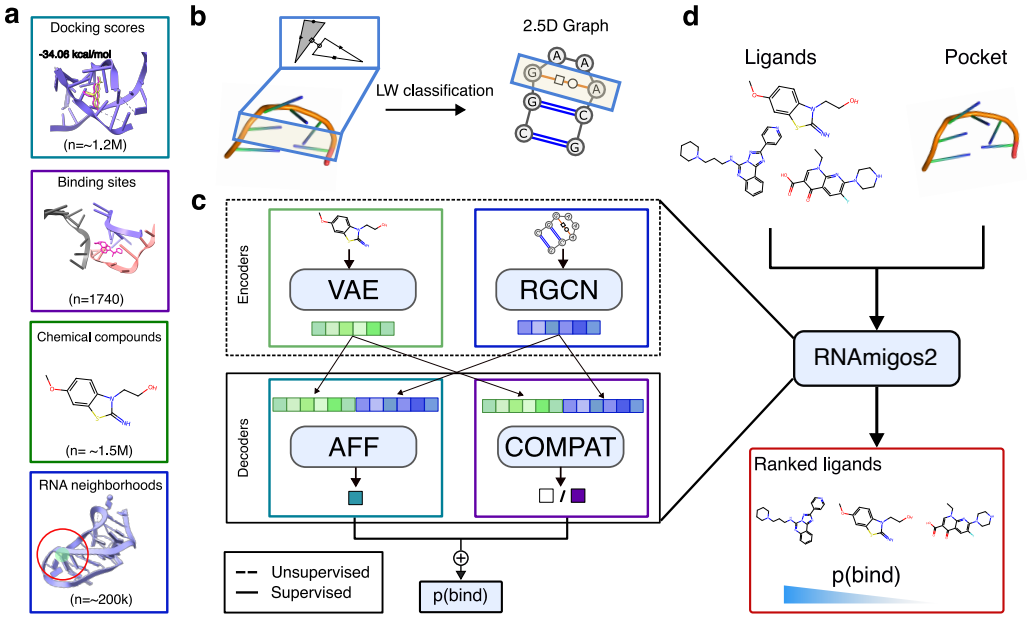


Figure 1: **a**: Datasets used in this paper. Beyond the supervised data available in the form of 1740 deposited structures of a ligand and its binding site, we use 1.2M molecular docking values as synthetic data and unlabeled datasets of 1.5M chemical compounds and 200k RNA residues neighborhoods. **b**: Representation of RNA structure as a 2.5D graph. Each binding site is encoded using Leontis-Westhof classification [27] and converted into an edge-attributed directed graph. **c**: Models to learn RNA-ligands interaction. We use an RNA and a ligand encoder, pre-trained in an unsupervised way. We then decode these representations to predict binding affinity (`Aff`) trained using our synthetic data; or a compatibility score (`Compat`) trained using deposited complexes. **d**: `RNAmigos2` integrate these models to screen input ligands libraries, sorting them by their predicted probability of binding.

`RNAmigos2` is designed to enable rapid screening of ligand libraries for binders using a query RNA structure. Our pipeline, illustrated in Figure 1, takes as input a candidate binding site structure (as full 3D or base pairing network) and a list of compounds to screen. The tool then returns a score for each compound reflecting binding likelihood.

`RNAmigos2` model uses an encoder-decoder framework, with two encoders and two decoders, which are each trained on separate data sources. The two encoders are respectively mapping the input RNA binding sites and small molecules into embeddings. The RNA 3D structure is represented as a graph, called 2.5D graph, that encodes all canonical (Watson-Crick and Wobble) and non-canonical base pair interactions occurring in the structure [28]. This representation allows us to capture key features of the RNA 3D architecture with a discrete mathematical object [27, 26] that suits well machine learning

frameworks and was shown to be a useful biological prior for RNA cheminformatics applications [37]. The RNA encoder takes the 2.5D graph as input and learns to generate an RNA representation using self-supervised training schemes [38] on all available non redundant RNA substructures. Ligands are represented as molecular graphs. The ligand encoder learns a neural representation for ligands, using a variational autoencoder model proposed in [8] and trained on large dataset of chemical compounds [20].

To train decoders, we extract 1,740 RNA-ligand complexes from the PDB[6] and group them in 436 clusters of similar binding sites we identified using RMAlign [70] with a similarity threshold of 0.75. Our first decoder (`Compat`) is trained as a binary classifier to distinguish between the native ligand of a binding site and decoys. Then, to synthetically augment the limited number and drug-likeness of PDB compounds, we perform a large-scale docking experiment by docking 500 drug-like ChEMBl chemical compounds on our 1,740 binding sites. Finally, our second decoder is trained to predict binding affinity (`Aff`) using the docking data. Given a binding site and a list of ligands, we encode all objects and use our joint decoders to predict a compatibility score that can be used for virtual screening. In what follows, we measure performance of a model by its ability to assign a high score to active compounds against a pool of inactive (decoy) compounds.

`RNAmigos2` is competitive with docking software and generalizes to new targets

In this work, we propose several enhancements to the original `RNAmigos` encoder model: we expand the dataset and training regimes, represent binding sites with directed graphs, and update the model architecture and pretraining strategy. Following the validation protocol in [37] and our ligand filtering, we show that the proposed enhancements yield a significant performance gain of over 25% AuROC points, achieving a new state-of-the-art on this task. The detailed results are presented in supplementary section E.

We now turn to an evaluation of our results using a larger and drug-like decoy set. The new decoy set is built querying the ChEMBL database [16] such that: (i) at most 100 compounds have a physico-chemical similarity ≥ 0.6 with the native ligand for each binding, (ii) the set has a high diversity detrmined by the MinMaxPicker algorithm ([24]), and (iii) all compounds pass a drug-like filter, keeping compounds with a molecular weight below 400, at least one ring, less than five rotatable bonds and H-bond donors, less than ten H-bond acceptors and logp values below five [49]. The active compound is defined as the one found co-crystallized in the PDB with each binding site, also termed the *native* ligand. On the binding site side, we enforce non-redundancy by computing the 3D alignment-based RMscore[70] over all pairs of binding sites in our dataset and perform a hierarchical agglomerative clustering with a cutoff of 0.75, resulting in 436 clusters of binding sites. The resulting binding site groups represent similar binding sites or different instances of the same binding site with potentially different ligands. Then, we performed a data splitting on the groups, ensuring no structural data leakage exists between our training set of 367 group of binding sites and our test set of 69 binding sites.

On this new dataset, we train models using our two settings: `Aff` and `Compat`, comparing performance to rDock on the test set. We find that `Aff` does indeed correlate well with rDock intermolecular energy terms (Spearman correlation of 0.73) on unseen binding sites, as well as shows a clear separation between native ligands and decoys with respect to both docking and `Aff` scores (see Fig. 2 a). Applying our models in a VS setting (Fig. 2**c**) we achieve a mean AuROC score of 0.958 and 0.939 using `Compat` and `Aff` models respectively. We emphasize the high performance of the `Aff` model that was trained solely on simulated data. It is worth noting that such high values are common in VS, as libraries of hundreds of thousands of compounds are screened but only a few thousands of compounds in the top percentiles can be sent to wet-lab assays. On the same task, rDock slightly outperformed our models with a mean value of 0.959. This result is expected given the intensive (and CPU-time costly) optimization process executed by rDock in identifying the best poses. We underline however that the runtime of our method is around ten seconds on a single machine, whereas the docking results were obtained after $\simeq 8$ CPU hours per binding site.

Interestingly, we find that the errors in VS aligned for each binding site for our different models in Figure 2**d**. While as expected, rDock and `Aff` errors align, the `Compat` trained on an independent data source has complementary errors modes. This suggests a simple strategy for mitigating their individual errors, and we propose to ensemble each model's results to create a model coined as `Mixed`. The `Mixed` scores now have a lower correlation of 0.68 with docking values (Figure 2**b**), but we

see that this correction enables to effectively recover active compounds that the docking program was not able to identify alone. We compare the VS performance of this mixing with ensemble of two models trained on the same data source, and report AuROCs over 3 independent seeds for all pairwise combinations in supplementary Table 7. We reach a performance of 0.990 AuROC for the `Mixed` to be compared to 0.965 and 0.940 for the ensembled versions of our single modality models. This performance gap proves the synergistic effect of our models, and enables us to significantly outperform rDock (0.959). Hence, the performance boost obtained by correcting docking scores with experimental data exceeds the decrease resulting from our imperfect docking surrogate model, which establishes a new state-of-the-art and best performing virtual screening tool.

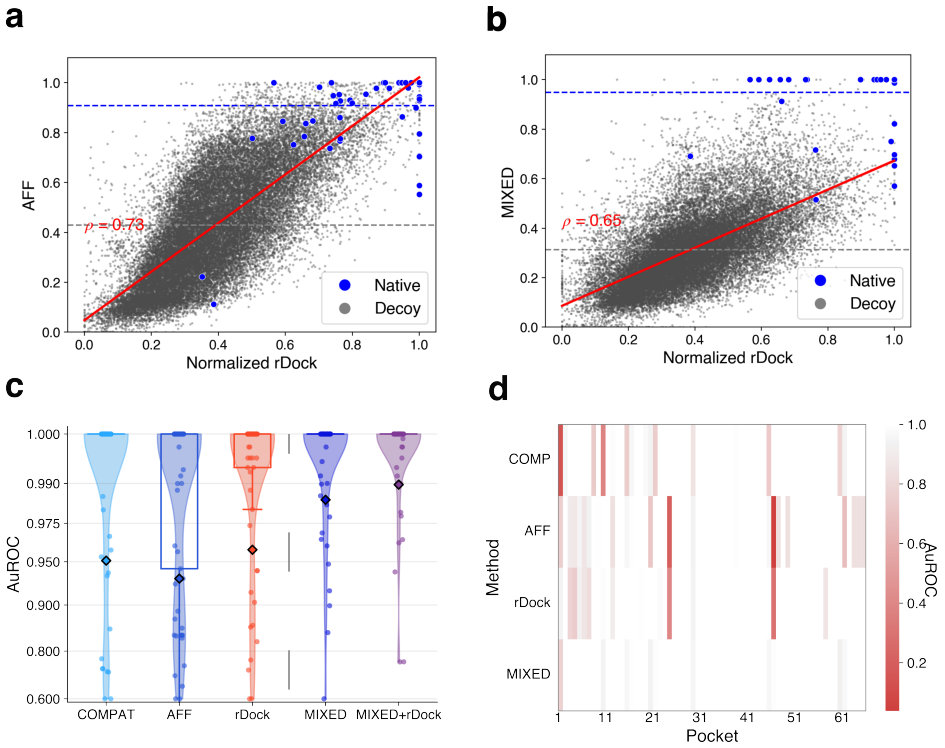


Figure 2: **a,b** Relationship between `Aff` (a) and `Mixed` predictions against rDock scores for decoys and native ligands in the test set. Horizontal lines indicate the mean of predicted scores for native and decoy ligands. **c** Distribution of the AuROC performance of different virtual screening models over test binding sites. Diamonds represent the mean performance. We compare models trained under our two settings with rDock (left of the dashed lines), an ensemble of these models (`Mixed`) and the average of `Mixed` and rDock. **d**. Description of the performance of the tools split by binding sites. Each line on this plot corresponds to a point in A. This shows that the successful predictions of different methods are complementary and illustrates their synergy in a mixed model.

Next, we coin our best performing configuration, `Mixed` as `RNAmigos2` for simplicity and evaluate it in a larger benchmark of RNA virtual screening tools and study the generalizability of our predictions. We compare the resulting model with several other VS tools, namely RLDOCK[53], AutoDock+Vina[58], Annapurna [51], rDock [46] and RNAmigos1[37]. As presented in Figure 3**a**, our method outperforms all methods with a wide margin for all methods and a smaller margin for rDock as seen in Fig 2**c**. We report minimal variance in our models' performances by training them over three random seeds and include these values in the first column of supplementary Table 7. Next, we look at the generalization capacity of `RNAmigos2` when considering the novelty of test set pockets and ligands with respect to the training data. For each binding site in the testing set we report the AuROC versus the structural similarity (RM-score [70]) to the most similar binding site in the training set (see Fig. 3b. We see that the model performs consistently well at ranges of structural redundancy from 0.3 up to the threshold value of 0.75.

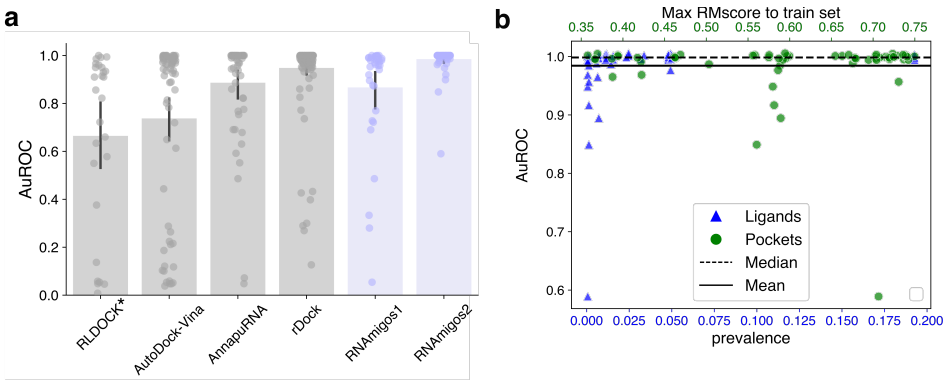


Figure 3: **a** Virtual screening benchmark against state of the art docking tools, RNAmigos1 and RNAmigos2. Note, RLDOCK produced errors on 42 binding sites. We only report its performance for 27 binding sites, which we denote with RLDOCK*. **b**. Depiction of the performance of `RNAmigos2` split by similarity of the test binding site to the train set. This similarity is assessed with RMScores for the binding sites, and Tanimoto similarity of chemical fingerprints for ligands.

`RNAmigos2` predictions consistently boost virtual screening efficiency

Our model outperforms docking by using complementary sources of information. However, the docking surrogate model alone (`Aff`) achieves a lower AuROC of 93.9% compared to 95.9% for actual docking. Thus, we also investigate mixing the `Compat` model with docking instead of `Aff`, and observe it enhances performance from 99.0% to 99.4%. We refer to this more accurate, but more expansive strategy, as `RNAmigos++`.

In practice, we have a limited time budget, and hence cannot afford to scan large scale databases with docking methods. This introduces a trade-off in virtual screening between quality of the scoring and the amount of computational time expended. When scanning a binding site, tool needs about 30 CPU seconds to perform the virtual screening, while the rDock model needs about 8 hours. Hence, in a drug design setting and under a compute budget of one CPU day, docking enables the screening of about $\sim$ 1400 compounds while `RNAmigos2` enables screening $\sim$ **1.4M compounds**.

To leverage the performance gain obtained with `RNAmigos++` while keeping computations tractable, we propose to quickly presort compounds according to our `RNAmigos2` score, and then sort the best scoring ones again with `RNAmigos++` (See Figure 4a). We expect that presorting the compounds will help the docking tool to review the most promising ligands first and thus use the allocated compute time more efficiently. We compare this hybrid strategy with the sole use of docking and with `RNAmigos2` in Figure 4b. Here, we define the *efficiency* of a VS method, as the area between the AuROC curve through time for `RNAmigos2` and rDock (further developped in Methods). The *efficiency* of our hybrid method is split by binding sites and plotted in Figure 4C.

This benchmark highlights that our methods obtain excellent results much faster than with a docking-only approach. We note that we benefit of a boost of performance with `RNAmigos++` over the `RNAmigos2` model after screening only the top compounds, which yields a ten-fold speedup. Finally, we show that we improve the performance for almost all binding sites, which suggests that our protocol never results in a waste of time (See Figure 4B). These results suggest that our method could allow the screening of larger libraries, potentially those coming from ligand generative models [7].

`RNAmigos2` identifies riboswitch ligands in large-scale *in vitro* assay

We now aim to assess the performance of `RNAmigos2` in a realistic large-scale scenario with *in vitro* validation. In the previous experiments, we established ground truth using co-crystallized ligands from the PDB as the active (native) compounds. Here, we use a recent large-scale *in vitro* screen of approximately 20k compounds against nucleic acid targets, known as ROBIN [68]. To make predictions using `RNAmigos2`, we identified all RNA targets in the ROBIN database which had a perfect sequence identity to an RNA chain in the PDB and pass the corresponding 3D structure as input to the model. Importantly, we ensure that all RNAs collected were not present in our training

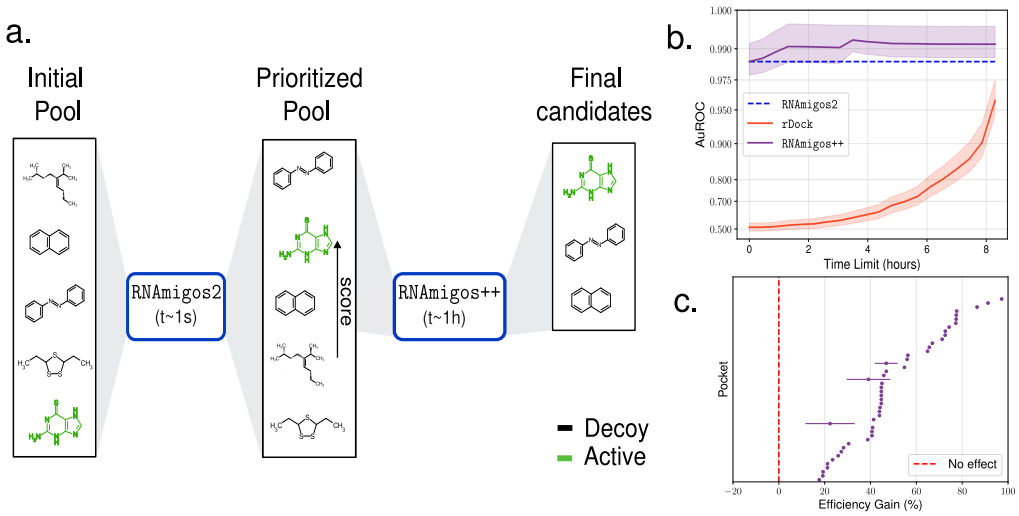


Figure 4: a. Illustration of our hybrid strategy. First, we sort the compounds using RNAmigos2 with a high throughput. Then, rDock is ran only on the best scoring compounds and used to replace our docking surrogate score. b. AuROC of different models, averaged over binding sites, as a function of time budget spent. c. Efficiency metric split by binding sites.

or validation sets. As RNAmigos2 requires a binding site as input, we use the native ligand in the PDB as before to define the input region. This results in four binding sites: TPP (PDBID: 2GDI), ZTP (PDBID: 5BTP), SAM_ll (PDBID: 2QWY), PreQ1 (PDBID: 3FU2) riboswitches, visualized in Fig.5a.

We computed the RNAmigos2 scores of all 24,572 assayed molecules in about four CPU minutes. We plot the chemical space induced by ROBIN compounds in Figure5a. Our top predictions (best 5%) are shown as colored full dots, with greater size for actives. Notably, we see that the retrieved active compounds span a large fraction of the chemical space. Nevertheless, we also note that some regions appear to be missed by our tool, such as the rightmost part of the plot. When plotting sample compounds from this region, we noted that the molecules seem to display sulfur atoms, which also decreases their drug likeness.

We plot the score distributions of actives (colored) and inactive compounds (grey) in Figure5c. This data again suggests that RNAmigos2 effectively differentiates actives from inactive compounds. We see that our tool is able to achieve enrichment factors at 1% ranging from 8.78 to 16.82, for an average of 12.2. Motivated by papers pointing out the low specificity of drug-target interaction predictor in proteins [59], we add a control experiment (RNAmigos2-Control Fig. 5c) where the prediction for each ligand is swapped between two pockets, eliminating the structure-specific signal while preserving the general affinity distribution. We observe positive enrichment factors (EF 1% 2.91-3.53) that illustrate how certain ligands seem to exhibit generic RNA binding properties. However, these enrichments are much weaker in the control setting, supporting the target-specificity of our predictions. Finally, we also include results obtained using rDock alone, and observe lower enrichments than when using RNAmigos2 (EF 1% 1.76-6.17). To our knowledge this is the first experimental evaluation of a deep learning structure-based model for RNA virtual screening.

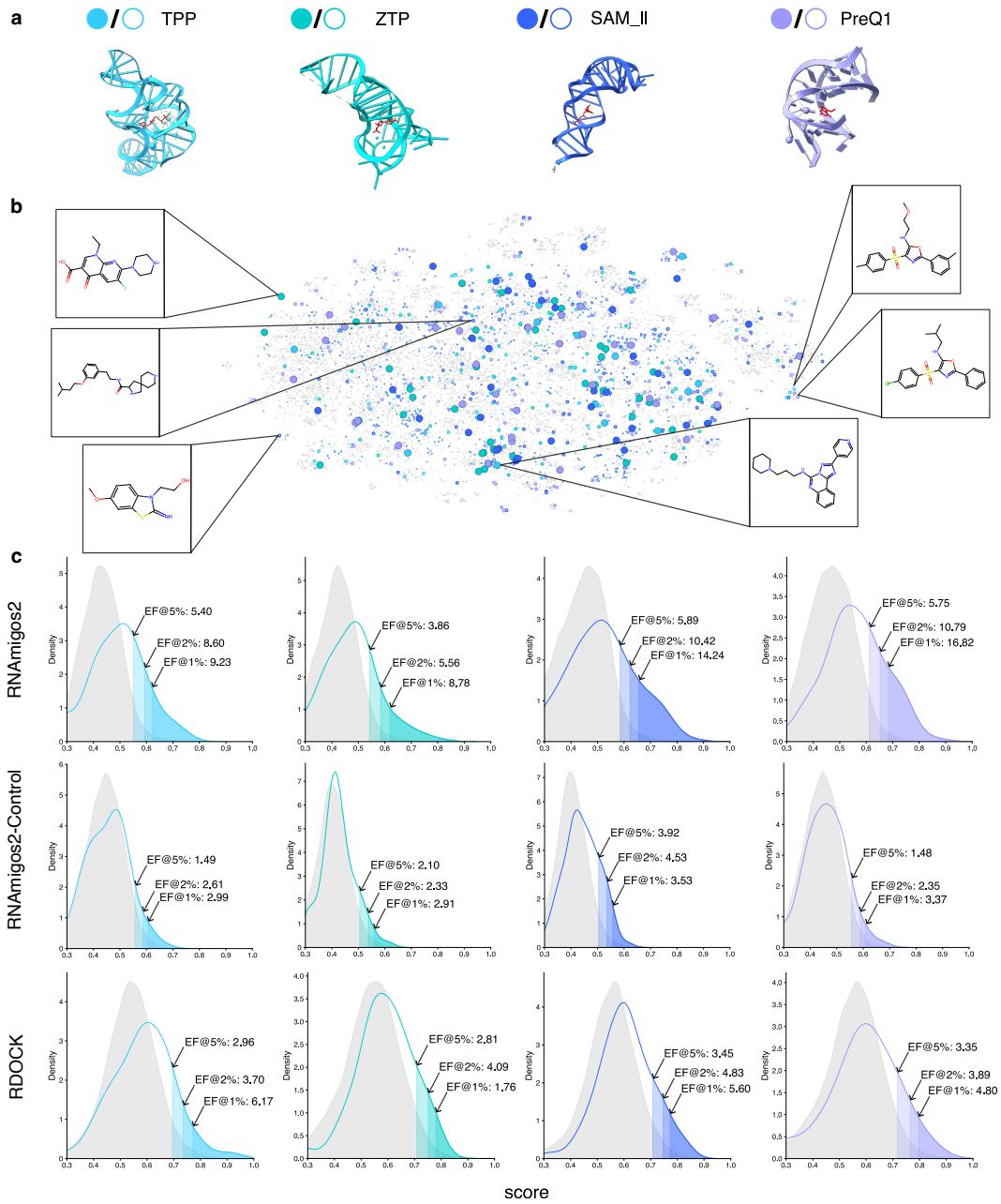


Figure 5: **a** Visualization of the structure of the riboswitches in the ROBIN dataset used in this experiment. **b** Chemical compounds selection (best 5%) in the tSNE space of ROBIN compounds. Large, colored compounds correspond to selected actives and smaller ones to selected inactive compounds. Hollowed ones are other actives and small grey points correspond to unselected, inactive compounds. are shown in color, with greater size for actives. **c** score distributions of actives (curve colored by pocket) and inactive compounds (grey curve) for each binding site. The shaded regions correspond to selecting the 5%, 2% and 1% top scoring compounds respectively. Score distributions obtained by shuffling predictions and using rDock are also provided.

Discussion

In this work, we address the timely question of *in-silico* RNA drug design. In contrast to proteins for which we already accumulated a lot of structural data, the current sparsity of reliable three-dimensional RNA structures prevents the application of machine learning paradigms for RNA drug discovery tasks. The development of dedicated frameworks and methods are therefore essential to address this bottleneck.

Specifically, we complement the available RNA-ligand complexes structures by using RNA docking as a data augmentation strategy and incorporate domain-specific unsupervised pre-training techniques. Whilst our docking surrogate model slightly under-performs traditional docking, its combination with a model trained on experimental data drastically outperforms three docking tools (0.96 rDock vs 0.99 `RNAmigos2` AuROC), in a fraction of runtime, for diverse targets. Moreover, by replacing the surrogate with actual docking scores for the top-scoring compounds, we manage to again half this error while maintaining a tenfold speedup over docking. Finally, we establish our tool's performance on an independent large-scale *in vitro* binding screen, and show that it provides twelve-fold enrichment factors in four minutes. Together, these results establish `RNAmigos2` as the new state of the art in structure-based RNA virtual screening.

By publicly releasing all our datasets, source code and model weights, we hope to spark a community effort in this important direction. Limitations to our approach include the need for pre-defined binding sites for which integrations with binding site predictors will need to be developed [60, 52, 53], as well as for the modeling of binding site flexibility [34]. We envision tools such as `RNAmigos2` will play a synergistic role with rapidly emerging RNA-centric technologies for molecular design [21] and newly released AlphaFold3 [1] which supports nucleic acids to pave the way for the next generation of RNA drug discovery. Notably, our methods have the distinct advantage to enable structure-based virtual screening for RNA with only low-resolution structural data at hand such as base pair interactions. In light of the daunting number of potential RNA targets, such feature could prove itself a major asset to mine entire genomes and fully embrace the era of RNA therapeutics [36].

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[2] Fardokht A Abulwerdi, Wenbo Xu, Abeer A Ageeli, Michael J Yonkunas, Gayatri Arun, Hyeyeon Nam, John S Schneekloth Jr, Theodore Kwaku Dayie, David Spector, Nathan Baird, et al. Selective small-molecule targeting of a triple helix encoded by the long noncoding rna, malat1. *ACS chemical biology*, 14(2):223–235, 2019.

[3] Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.

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

[5] Florent Barbault, Liangren Zhang, Lihe Zhang, and Bo Tao Fan. Parametrization of a specific free energy function for automated docking against rna targets using neural networks. *Chemometrics and Intelligent Laboratory Systems*, 82:269–275, 2006.

[6] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[7] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[8] Jacques Boitreaud, Vincent Mallet, Carlos Oliver, and Jerome Waldispuhl. Optimol: optimization of binding affinities in chemical space for drug discovery. *Journal of Chemical Information and Modeling*, 60(12):5658–5666, 2020.

[9] Adrià Cereto-Massagué, Laura Guasch, Cristina Valls, Miquel Mulero, Gerard Pujadas, and Santiago Garcia-Vallvé. Decoyfinder: an easy-to-use python gui application for building target-specific decoy sets. *Bioinformatics*, 28(12):1661–1662, 2012.

[10] Hongming Chen, Paul D Lyne, Fabrizio Giordanetto, Timothy Lovell, and Jin Li. On evaluating molecular-docking methods for pose prediction and enrichment factors. *Journal of chemical information and modeling*, 46(1):401–415, 2006.

[11] Matthew D Disney. Targeting rna with small molecules to capture opportunities at the intersection of chemistry, biology, and medicine. *Journal of the American Chemical Society*, 141(17):6776–6790, 2019.

[12] Matthew D Disney, Audrey M Winkelsas, Sai Pradeep Velagapudi, Mark Southern, Mohammad Fallahi, and Jessica L Childs-Disney. Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured rnas. *ACS chemical biology*, 11(6):1720–1728, 2016.

[13] Matthew D Disney, Audrey M Winkelsas, Sai Pradeep Velagapudi, Mark Southern, Mohammad Fallahi, and Jessica L Childs-Disney. Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured rnas. *ACS chemical biology*, 11(6):1720–1728, 2016.

[14] James P Falese, Anita Donlic, and Amanda E Hargrove. Targeting rna with small molecules: from fundamental principles towards the clinic. *Chemical Society Reviews*, 50(4):2224–2243, 2021.

[15] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mrnas are conserved targets of microrrnas. *Genome research*, 19(1):92–105, 2009.

[16] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[17] Oleksandr O Grygorenko, Dmytro S Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E Gubina, and Yurii S Moroz. Generating multibillion chemical space of readily accessible screening compounds. *Iscience*, 23(11), 2020.

[18] Christopher L Haga and Donald G Phinney. Strategies for targeting rna with small molecule drugs. *Expert Opinion on Drug Discovery*, 18(2):135–147, 2023.

[19] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

[20] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[21] Chaitanya K Joshi, Arian R Jamasb, Ramon Viñas, Charles Harris, Simon V Mathis, Alex Morehead, and Pietro Liò. grnade: Geometric deep learning for 3d rna inverse design. *bioRxiv*, pages 2024–03, 2024.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[23] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[24] Greg Landrum, Paolo Tosco, Brian Kelley, Ric, David Cosgrove, Sriniker, Gedeck, Riccardo Vianello, NadineSchneider, Eisuke Kawashima, Dan N, Gareth Jones, Andrew Dalke, Brian Cole, Matt Swain, Samo Turk, AlexanderSavelyev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Daniel Probst, Kazuya Ujihara, Vincent F. Scalfani, Guillaume Godin, Juuso Lehtivarjo, Rachel Walker, Axel Pahl, Francois Berenger, Jasondbiggs, and Strets123. rdkit/rdkit: 2023 03 3 (q1 2023) release, 2023.

[25] Antonio Lavecchia and Carmen Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry*, 20(23):2839–2860, 2013.

[26] 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. Nucleic acids/Sequences and topology.

[27] NEOCLES B. LEONTIS and ERIC WESTHOF. Conserved geometrical base-pairing patterns in rna. *Quarterly Reviews of Biophysics*, 31(4):399–455, 1998.

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

[29] Neocles B Leontis and Craig L Zirbel. Nonredundant 3d structure datasets for rna knowledge extraction and benchmarking. *RNA 3D structure analysis and prediction*, pages 281–298, 2012.

[30] Ding Luo, Dandan Liu, Xiaoyang Qu, Lina Dong, and Binju Wang. Enhancing generalizability in protein–ligand binding affinity prediction with multimodal contrastive learning. *Journal of Chemical Information and Modeling*, 64(6):1892–1906, 2024.

[31] 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.

[32] Eduardo Habib Bechelane Maia, Letícia Cristina Assis, Tiago Alves De Oliveira, Alisson Marques Da Silva, and Alex Gutterres Taranto. Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry*, 8:343, 2020.

[33] Vincent Mallet, Carlos Oliver, Jonathan Broadbent, William L Hamilton, and Jérôme Waldispühl. Rnaglib: a python package for rna 2.5 d graphs. *Bioinformatics*, 38(5):1458–1459, 2022.

[34] Nicolas Moitessier, Eric Westhof, and Stephen Hanessian. Docking of aminoglycosides to hydrated and flexible rna. *Journal of medicinal chemistry*, 49(3):1023–1033, 2006.

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

[36] Kinga Nemeth, Recep Bayraktar, Manuela Ferracin, and George A Calin. Non-coding rnas in disease: from mechanisms to therapeutics. *Nature Reviews Genetics*, 25(3):211–232, 2024.

[37] 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*, 48:7690–7699, 2020.

[38] Carlos Oliver, Vincent Mallet, Pericles Philippopoulos, William L Hamilton, and Jérôme Waldispühl. Vernal: a tool for mining fuzzy network motifs in rna. *Bioinformatics*, 38(4):970–976, 2022.

[39] 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. *Bioinformatics*, 38(17):4185–4193, 2022.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[41] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[42] Anna Philips, Kaja Milanowska, Grzegorz Łach, and Janusz M Bujnicki. Ligandrna: computational predictor of rna–ligand interactions. *Rna*, 19(12):1605–1616, 2013.

[43] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Sergey I. Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *CoRR*, abs/1811.12823, 2018.

[44] Alexander S Powers, Helen H Yu, Patricia Suriana, Rohan V Koodli, Tianyu Lu, Joseph M Paggi, and Ron O Dror. Geometric deep learning for structure-based ligand design. *ACS Central Science*, 2023.

[45] Peter W Rose, Chunxiao Bi, Wolfgang F Bluhm, Cole H Christie, Dimitris Dimitropoulos, Shuchismita Dutta, Rachel K Green, David S Goodsell, Andreas Prlić, Martha Quesada, et al. The rcsb protein data bank: new resources for research and education. *Nucleic acids research*, 41(D1):D475–D482, 2012.

[46] Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, and S. David Morley. rdock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Computational Biology*, 10:1–8, 2014.

[47] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[48] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[49] Suliman Sharif, Ruibin Liu, Asuka A. Orr, Daniel Khavrutskii, Sunhwan Jo, Bettina Lier, Anastasia Croitoru, Chris Burke, Aziza Frank, Jacob Weiner, Nathaniel McClean, Aarion Romany, Mingtian Zhao, Takayuki Serizawa, Jared Deacon, Ian Jones, Shaoqi Zhan, Anmol Kumar, Mike Woster, Rebecca Pinette-Dorin, Elena Y. Chow, Sevien Schulhoff, and Alexander D. MacKerell Jr. Global-Chem: A Chemical Knowledge Graph of common small molecules and their IUPAC/SMILES/SMARTS for selection of compounds relevant to diverse chemical communities, August 2022.

[50] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pages 20503–20521. PMLR, 2022.

[51] Filip Stefaniak and Janusz M Bujnicki. Annapurna: A scoring function for predicting rna-small molecule binding poses. *PLoS computational biology*, 17(2):e1008309, 2021.

[52] Hong Su, Zhenling Peng, and Jianyi Yang. Recognition of small molecule–rna binding sites using rna sequence and structure. *Bioinformatics*, 37(1):36–42, 2021.

[53] 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.

[54] Saisai Sun and Lin Gao. Contrastive pre-training and 3d convolution neural network for rna and small molecule binding affinity prediction. *Bioinformatics*, 40(4):btae155, 2024.

[55] Natalia A Szulc, Zuzanna Mackiewicz, Janusz M Bujnicki, and Filip Stefaniak. Structural interaction fingerprints and machine learning for predicting and explaining binding of small molecule ligands to rna. *Briefings in Bioinformatics*, 24(4):bbad187, 2023.

[56] D Thirumalai and Changbong Hyeon. Rna and protein folding: common themes and variations. *Biochemistry*, 44(13):4957–4970, 2005.

[57] Shusuke Toden, Timothy J Zumwalt, and Ajay Goel. Non-coding rnas and potential therapeutic targeting in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1875(1):188491, 2021.

[58] 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*, pages NA–NA, 2009.

[59] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958, 2022.

[60] 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.

[61] Kaili Wang, Renyi Zhou, Yaohang Li, and Min Li. Deepdtaf: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, 22(5):bbab072, 2021.

[62] Minjie Yu Wang. Deep graph library: Towards efficient and scalable deep learning on graphs. In *ICLR workshop on representation learning on graphs and manifolds*, 2019.

[63] Renxiao Wang, ueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 22, 11 2005.

[64] K. D. Warner, C. E. Hajdin, and K. M. Weeks. Principles for targeting rna with drug-like small molecules. *Nature Reviews Drug Discovery*, 17:547–558, 2018.

[65] Leiming Xia, Lei Xu, Shourun Pan, Dongjiang Niu, Beiyi Zhang, and Zhen Li. Drug-target binding affinity prediction using message passing neural network and self supervised learning. *BMC genomics*, 24(1):557, 2023.

[66] Juan Xu, Kang-jing Wu, Qiao-jun Jia, and Xian-feng Ding. Roles of mirna and incrna in triple-negative breast cancer. *Journal of Zhejiang University-science b*, 21(9):673–689, 2020.

[67] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4, 2006.

[68] Kamyar Yazdani, Deondre Jordan, Mo Yang, Christopher R Fullenkamp, David R Calabrese, Robert Boer, Thomas Hilimire, Timothy EH Allen, Rabia T Khan, and John S Schneekloth Jr. Machine learning informs rna-binding chemical space. *Angewandte Chemie*, 135(11):e202211358, 2023.

[69] Yuqi Zhang, Stefano Forli, Anna Omelchenko, and Michel F. Sanner. Autogridfr: Improvements on autodock affinity maps and associated software tools. *Journal of Computational Chemistry*, 40:2882–2886, 2019.

[70] Jinfang Zheng, Juan Xie, Xu Hong, and Shiyong Liu. Rmalign: an rna structural alignment tool based on a novel scoring function rmscore. *BMC Genomics*, 20:1–10, 2019.

Methods

Constructing and representing a Dataset of Binding Sites

To construct our dataset of binding sites, we download all RNA structures from the Protein Data Bank (PDB)[45] containing RNA and at least one ligand. Then, we extract RNA binding sites as RNA residues with an atom at most 10Å from the ligand, and filter out resulting binding sites with less than five RNA residues or involving a majority of protein residues. We additionally remove complexes with ligands that did not follow rules introduced in Hariboss[39] (e.g. removing ions, crystallization artefacts, etc.). The final dataset consists of **1,740 binding sites** associated with **264 unique ligands**.

We then use RNAGlib[33] to obtain a representation of the structure of our binding sites in the form of 2.5D graphs. The nodes of these graphs represent nucleotides and contain a one-hot encoding of the nucleotide type, and the edges represent interactions between nucleotides, split in interaction types that follow the Leontis–Westhof classification[28]. An undirected 2.5D graph representation was shown to give better performance[37] than traditional representations for machine learning. In this work, we additionally use directed edges relevant to disambiguate sequence orientation and asymmetrical non canonical interactions. We expanded our graphs by performing a breadth-first expansion of depth four, to add context nodes around the binding sites.

Unsupervised Datasets of RNA structures and chemical compounds

We use two additional sources of unsupervised data to pretrain our networks. We use all RNA structures in the non redundant dataset proposed in [29], represented as directed 2.5D graphs, and extract the 2-hops neighborhoods around each residue, resulting in over 200,000 small RNA graphs. We additionally use a curated dataset of $\sim$1.5M ZINC[20] compounds, introduced in [43] as reasonable drug like compounds. These molecular compounds are represented as molecular graphs with nodes containing atom types, charge and aromaticity and edges with different chemical interactions.

Docking Dataset Construction

Our current database only includes co-crystallized ligands, that are thus positive binding examples. We do not have information about the affinity between a binding site and the native ligands of other binding sites, which are thus assumed to be negative pairs. Moreover, we would like to have more affinity data about drug-like compounds, since PDB compounds are not necessarily drugs (cofactors for instance). To generate more data, we will use molecular docking as a source of synthetic data.

Therefore, we curate a list of drug-like compounds from ChEMBL[16]. To do so, we use a drug-like filter and keep compounds with a molecular weight below 400, at least one ring, less than five rotatable bonds and H-bond donors, less than ten H-bond acceptors and logp values below five [49]. We complement this procedure with the MaxMin algorithm based on Tanimoto distance between fingerprints, using RDKit[24], for picking a subset of 500 diverse compounds, ensuring that up to 100 compounds have a Tanimoto Coefficient greater than 0.6 with the native ligand for each binding site.

To perform our molecular docking, we compared two candidate tools, rDock [46] and AutoDock Vina[58] which were identified in a recent review of nucleic-acid docking tools[31]. In addition, we used AnnapuRNA[51], a knowledge-based scoring function, as an independent referee to compare these candidate software in terms of native pose identification. We first establish our docking procedure and compare our candidate docking programs using self and cross docking experiments. In self-docking experiments, we evaluate programs' ability to predict the binding modes inside the native macromolecule by comparing the generated ligand poses to the crystal structure. In cross-docking experiments, we evaluate the ability of docking tools to retrieve actives among decoys. An in-depth explanation of these experiments is available in supplementary section C. Based on our results, we choose to use rDock as our docking program. We used it to dock the *pdb* and *chembl* ligand sets of 264 native binders and 500 ChemBL compounds in our 1740 binding sites, resulting in over 1.3 million values, that can be downloaded at the following link[2]. On the test set, the docking scores were used to obtaining virtual screening performance of rDock for benchmarking purpose. The ones obtained for the train set binding sites are used as a data augmentation strategy.

[2] `github.com/cgoliver/rnamigos2`

Data splitting

We compute the RMscore[70] of all pairs of binding sites in our dataset, and report their distribution in supplementary Figure 6a. As can be seen, while most of the binding sites are distincts, there exists a proportion of binding sites that are highly similar. Hence, we then perform a hierarchical agglomerative clustering with a cutoff of 0.75. This allows us to group similar binding sites together and results in **436 groups** of binding sites.

Those groups can represent similar binding sites, or different instances of the same binding site with different ligands. Data is split based on the groups, ensuring no structural data leakage exists between our training set of 367 group of binding sites and our test set of 69 binding sites. An MDS representation of our train and test groups is available in Figure 6b, showing that our test points are diverse in the space of binding sites. Training and testing on those groups requires caution. To avoid data imbalance, we chose to choose only a representative per group, and to set all ligands of this group as *group actives*. While training, we chose to dynamically sample one of the group actives. While testing and to compute AuROCs, all the group actives are used as positives.

Models

Our model adopts an encoder-decoder framework with two encoders and two decoders. Our encoders are pre-trained to serve as a strong initialization for the representation of the input binding sites and the small molecules. For the binding site encoding, we use the directed 2.5D graph representation. Given such a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we then use a neural network encoder f_θ based on Relational Graph Convolutional Network (RGCN)[47] layers to produce node embeddings of dimension $d \in \mathbb{N}$ and a pooling step reduces those node embeddings to a single graph-level vector representation $\phi_\mathcal{G} = f_\theta(\mathcal{G})$. Our RNA encoder using the strategies proposed in[38] which are based on the metric learning framework[67]. Namely, given a similarity metric k, the network is trained to minimize

$$\mathcal{L} = ||\langle \phi_{\mathcal{G}_1}, \phi_{\mathcal{G}_2} \rangle - k(\phi_{\mathcal{G}_1}, \phi_{\mathcal{G}_2})||^2.$$

Our directed graphs open the door to using more powerful similarity functions, and hence enhance pre-training. In this work we used a graphlet matching algorithm that matches directed rooted subgraphs.

To encode the structure of small molecules, the ligand is represented as a molecular graph $\mathcal{F}$ to serve as input to our model. This graph is then fed into a ligand encoder, denoted as g_θ, resulting in $\phi_\mathcal{F} = g_\theta(\mathcal{F})$. We use the architecture and pretraining scheme proposed in OptiMol[8]: we additionally encode the graph as a SELFIES [23] string and try to reconstruct this string from $\phi_\mathcal{F}$ following a Variational Auto-Encoder (VAE) framework.

We then concatenate the graphs and ligand embeddings and feed the result to a decoder that outputs real values, so that we can write our overall models as $M(\mathcal{G}, \mathcal{F}) = h_\theta(f_\theta(\mathcal{G}) \bigoplus g_\theta(\mathcal{F})) \in \mathbb{R}$. Our first decoder (`Compat`) uses this architecture and is trained to output one for native ligands and zero for decoys, by minimizing a Binary Cross Entropy (BCE) loss over experimental complexes Our second setting (`Aff`), keeps the M model described above, but changes the training procedure so that the inter-atomic term of the docking score becomes the prediction target. We normalize docking scores to be between zero and one and train our network to approximate this value using an $\mathcal{L}_2$ loss. The specific model architecture and hyper-parameter settings are described in supplementary Section D.

Inference and validation with virtual screening

Virtual screening (VS) aims to sort a library of compounds such that the active ones are at the top of the list[25]. Typically, a VS is used to screen millions of compounds, as only a few thousands can be sent to wet-lab assays. Metrics for VS quantify the gain in active compounds obtained by selecting compounds using a VS strategy instead of a random selection. The main VS metric are the well known AuROC and Enrichment Factors[10], where the enrichment factor at a cutoff X of a VS method is the fraction of actives versus decoys (non-actives) in the top scoring $X\%$ relative to the overall active fraction.

Because the number of drug-like compounds available in databases such as Enamine[17] is in the billions, we introduce a novel metric, denoted *efficiency*, to measure the expected performance of

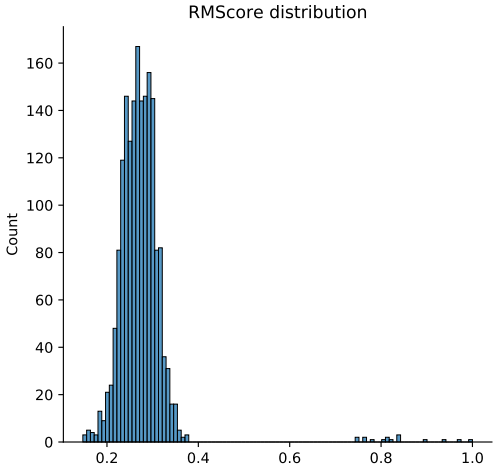
a screening after a certain time budget is spent. While performing a virtual screening over a list of compounds, we keep track of the AuROC as a function of time and define the *efficiency* of a search as the area under the time vs. AuROC curve. The value of this metric depends on the initial ordering of the list. For instance, if all active compounds are at the end of the list, they are processed last, yielding zero AuROC until these compounds are reached. To get an expected efficiency, we thus repeat the computation over many shuffled initial orderings . In this manner we can capture the trade-off between time expenditure and the quality of the search results in a single value.

To perform virtual screening with our tool, we simply compute the scores for all of our compounds and sort them based on this score. To avoid imbalance when mixing our models, we first normalize the distributions of scores independently for the `Compat` and the `Aff` and then take the average of the results as the predicted score. To use our tool in conjunction with docking, we quickly presort the compounds using our mixed score and only dock the prioritized compounds. For those compounds, we use the true docking score instead of their approximation by the `Aff`.

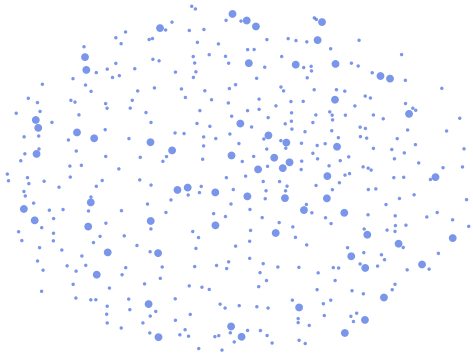
Appendix

A Properties of RNA binding sites

In this section, we analyze the properties of our binding sites and their representation as 2.5D graphs. First, we assess the redundancy of our binding sites. We use RMScore[70] to measure and analyze RNA structural similarity on our binding sites. The distribution of those values, shown in Figure6a), displays a strong peak below 0.3. We compute a multidimensional scaling [41] of our data based on this similarity measure in Figure 6b. In this representation, dissimilarity of the samples correlates with the distance between points on the plot. The absence of clusters in this representation suggests that the data-set is not redundant. When trying to enforce a clustering using RMscore values, the silhouette score has a value below 0.3, showing that the clusters are bad and hence, that sites are not similar.



(a) RMScore distribution over binding pockets.



(b) 2D MDS projection of pockets. Pockets in the test sets are shown as larger points

Figure 6: Pocket structural similarity analysis.

We now turn to an analysis of the representation of our binding sites. Table 1 reports the main features of the graphs we extracted for our 1740 binding sites.

	Nodes	Edges	Non-canonicals	Canonicals	Connected components
Mean	10.3	11.6	1.47	10.1	1.20
Std	4.0	5.9	2.1	5.3	0.42
Min.	4	2	0	0	1
25%	7	6	0	5	1
50%	11	12	1	10	1
75%	13	16	2	15	1
Max.	22	33	18	23	3

Table 1: Descriptive statistics of the binding sites.

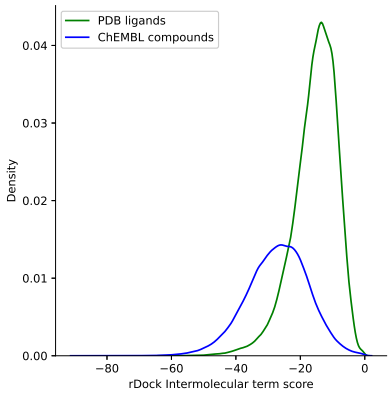
B Ligand properties

Next, we turn to analyzing the ligands used in this paper. We compare properties of our two ligand sets. As a reminder, our first set is 264 ligands filtered from the PDB, used to compare to RNAmigos and to train our `Compat` model. Our second set is made of 500 ligands per binding site extracted from

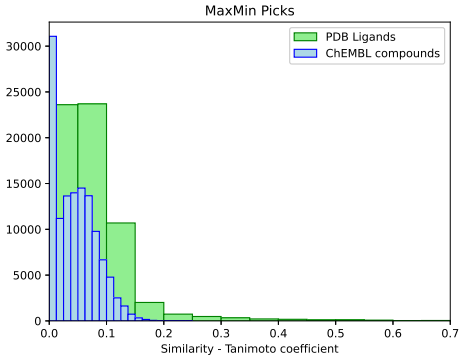
the ChemBL database and selected for diversity and drug-likeness. We present a few properties of those compounds in Table 2. Moreover, we compute the distribution of inter-molecular energies for the ligands (Figure 7a) and the distribution of pairwise similarities (Figure 7b).

Descriptor	PDB	ChEMBL
Molecular weight	455.5	271.9
Logp	-0.083	1.315
H bond donor	4.467	1.776
H bond acceptors	9.05	4.6
Rotatable bonds	6.3	2.3
Number of atoms	31.6	19.2
Molar refractivity	113	71.7
Topological surface area mapping	159	74
Formal charge	0.20	0.04
Heavy atoms	31.6	19.3
Number of rings	3.2	2.9

Table 2: Chemical descriptors of PDB Ligands and ChEMBL compounds



(a) Rdock score distribution for the PDB and ChEMBL datasets.



(b) Distribution of pairwise distances for the PDB and ChEMBL datasets.

We can see that the PDB ligands do not display drug-like features. In particular, the compounds are bigger than drug-like compounds. They also display worse docking scores on average. These results indicate that this second ChEMBL was indeed needed for an assessment of our tool from a drug design perspective. Finally, our results show that the samples in each of the data sets are diverse.

Finally, in Figure 8 we plot a 2D MDS of our ligand space in which we identify test compounds. This illustrates that our actives span a large fraction of the chemical space.

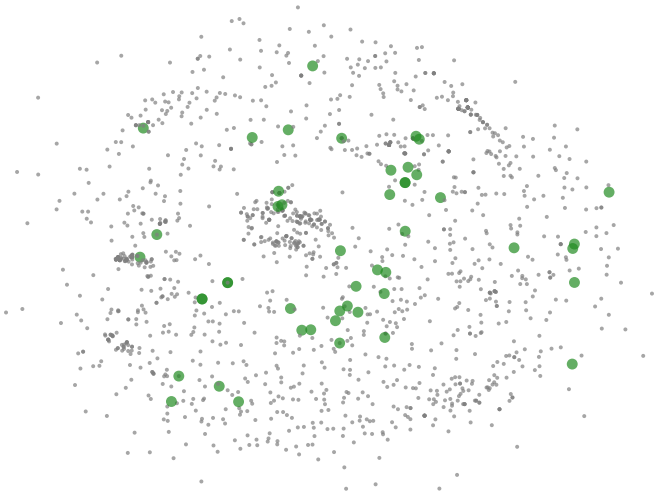


Figure 8: RMScore distribution over binding pockets.

C Docking software benchmark

We assessed two candidate docking tools, rDock [46] and AutoDock Vina[58] which were identified in a recent review of nucleic-acid docking tools[31]. In addition, we used AnnapuRNA[51], a knowledge-based scoring function, as an independent referee to compare these candidate software in terms of native pose identification. We first establish our docking procedure and compare our candidate docking programs using self and cross docking experiments. In self-docking experiments, we evaluate programs' ability to predict the binding modes inside the native macromolecule by comparing the generated ligand poses to the crystal structure. In cross-docking experiments, we evaluate the ability of docking tools to retrieve actives among decoys. Based on both experiments, we choose to use rDock as our docking program. An in-depth explanation of these experiments is available in section C.

Let us describe the settings with which we have run our docking experiments, using AutoDock Vina and rDock. For AutoDock Vina, we defined the exhaustiveness, energy_range and num_modes to be 8, 3 and 100, respectively. The search space (i.e., the coordinates for the center of mass and the size of the binding sites) was calculated by AutoGridFR [69]. We also executed a different set of experiments by using the parameters presented in [46] (i.e., exhaustiveness=16, num_modes=100, energy_range=30 and the search space calculated by rDock). The greatest affinity difference between (the best pose predicted by) the two sets of parameters was 0.3 kcal/mol. Regarding the RMSD score, the greatest difference occurred for the 1BYJ structure (5.404); however, all of the other structures presented differences lower than 0.183. These results agree with the developers of AutoDock Vina who stated that *"With the default (or any given) setting of exhaustiveness, the time spent on the search is already varied heuristically depending on the number of atoms, flexibility, etc. Normally, it does not make sense to spend extra time searching to reduce the probability of not finding the global minimum of the scoring function beyond what is significantly lower than the probability that the minimum is far from the native conformation"*.

Once a system is chosen, rDock performs a cavity generation first, and then the actual docking. As in [46], the cavity was defined using the crystallographic ligand (found in the PDB data-base) as reference through the "reference ligand method" implemented in rDock. All the parameters were used by default; but the radius, small_sphere, max_cavities, flexibility were set to 4.0, 1.0, 1 and 3, respectively. The models were scored using the "dock_solv.prm" function.

Self and cross docking experiments are performed to have a fair comparison between the capabilities of the selected tools when used on RNA molecules as a target.

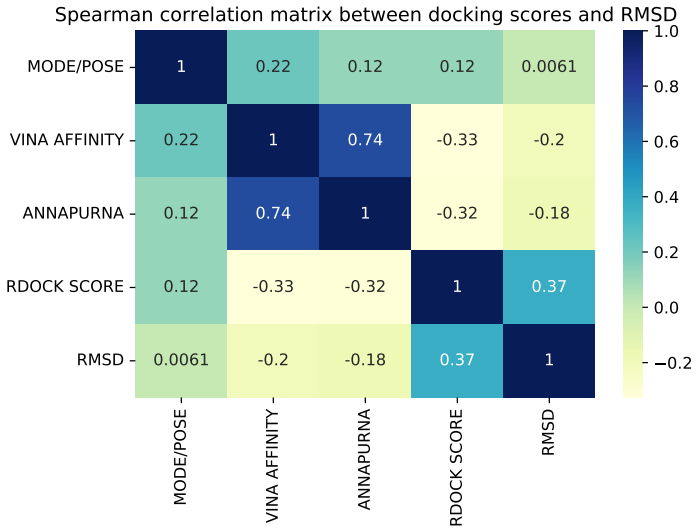


Figure 9: Correlation matrix between docking scores and RMSD to native pose.

C.1 Self-docking experiments

During the self-docking experiments, we evaluate the ability of the docking tools to predict the binding modes inside the native macromolecule. Therefore, we compare the generated ligand poses predicted as best to the crystal structure. We use the root mean square deviation (RMSD) as the metric to determine the accuracy of the prediction.

The experiments were run on a set of 9 RNA-ligand complexes (1EI2, 1QD3, 1LVJ, 1BYJ, 1NEM, 1PBR, 1TOB, 2TOB, and 1FMN) obtained from [5] and [69]. For AutoDock Vina, the complex structures were downloaded from the PDB data-base. For rDock, the structures were obtained from the validation dataset reported by [46].

Table 3 summarizes the descriptive statistics for the RMSD of the 20 best poses for each complex and for each tool. The total number of RSMD values for each docking tool is 180 (twenty for each complex). The mean and median RMSD are more favorable for rDock with 4.28A and 3.62A, respectively (7.21A and 7.43 for AutoDock Vina). The distribution of RMSD values is shown in Figure 9A. This figure confirms that, in general, the RMSD values for the poses predicted by rDock are lower than those predicted by AutoDock Vina.

Docking tool	Mean	Std.	Min.	25%	50%	75%	Max.
rDock	4.28	2.85	0.68	1.84	3.62	6.78	10.32
AutoDock Vina	7.21	2.02	1.17	5.79	7.43	8.84	10.74

Table 3: Descriptive statistics of the RMSD obtained with AutoDock Vina and rDock

Figure 9B depicts the correlation matrix between the rDock scores, AutoDock Vina affinities, AnnapuRNA scores, and RMSD values. There is a positive high correlation (0.74) between AutoDock Vina and AnnapuRNA scoring functions. Regarding rDock, there are weaker negative correlations with AutoDock Vina (-0.33) and AnnapuRNA (-0.32); however, there is a moderate positive correlation (0.37) with the RMSD values. Considering the RMSD values, its correlations with respect to AutoDock Vina and AnnapuRNA scores are -0.2 and -0.18, respectively.

Annapurna can be used to re-rank the generated poses. If we take the top 5 poses as ranked by AnnapuRNA, the mean RMSD decreases in 6 out of 9 complexes (for both AutoDock Vina and rDock). If only the top pose is considered, 3 complexes and 5 complexes are improved for AutoDock and rDock, respectively.

Based on the self-docking experiment, we can conclude that rDock has better performance than AutoDock Vina in docking ligands against RNA targets. In addition, re-scoring the predicted poses with AnnapuRNA identifies poses with lower RMSD values than those originally ranked by AutoDock Vina and rDock.

C.2 Cross-docking experiments

In cross docking experiments, we evaluate the ability of docking tools to retrieve actives among decoys. We do so by docking each ligand in the dataset to each RNA target. The experiments were carried out on a set of 52 RNA-ligand complexes, having 31 unique ligands. We use the native ligand of each complex as the active ligand and the other 30 ligands in the dataset as inactive ligands.

AutoDock Vina only returns the binding affinity of the bound conformations calculated with its scoring function, and on the other hand, in addition to the total score, rDock details the value of each term of the scoring function (intermolecular, ligand intramolecular, site intramolecular and external restraint). Beside the total score, we also ranked the poses predicted with rDock based on the intermolecular term because it represents the RNA-ligand interaction score [46]. We use AnnapuRNA to re-score the poses. The results are presented in Table 4 each docking tool and for the poses. We find rDock with the intermolecular term of its scoring function to have the best performance. Based on the results shown in this section, we chose rDock as our docking tool.

	Vina	rDock Total	rDock Inter.	rDock AnnapuRNA	Vina AnnapuRNA
AuROC	0.64	0.69	0.73	0.61	0.60

Table 4: Mean and standard deviation of the AuROC scores obtained with different settings. We compare the use of the Rdock score, the use of only its intermolecular term, Rdock with Annapurna and AutoDock Vina with Annapurna.

C.2.1 Runtime for docking experiments

Table 5 reports the execution time it took to perform the docking experiments, including the cavity generation stage and the docking itself. The method selected for generating the cavity was to take the native ligand as the reference ligand.

Cavity generation		Docking - ChEMBL	Docking - PDB
# Pockets	Time (minutes)	Avg time per compound (min.)	Avg time per compound (min.)
645	5		
85	8		
335	20		
302	40	0.41	1.5
210	60		
35	120		

Table 5: Docking experiments runtimes.

The approach rDock uses to parallelize the jobs is independently running each molecule, splitting the ligands file into multiple files, and running each independently. To parallelize the docking experiments, we executed the experiments over each pocket in parallel using an array job in Compute Canada. Therefore, for the cavity generation task we used as many CPU numbers as pockets (1740), taking approximately 2 hours to complete this task. In the same sense, the docking of each pocket with the ChEMBL compounds, using 1740 CPUs, is approximately 4 hours and for the PDB ligands, 7 hours.

D Model Architecture and Training

The RNA encoder f_θ is implemented as a Relational Graph Convolutional Network (RGCN) [48] with 3 convolutional layers, and a hidden dimension of 64 and no basis sharing. The pretraining of this model is done using the RNAGlib package [33] with default parameters. The resulting node embeddings are pooled into a graph-level embedding $\phi_{\mathcal{G}}$ with a global attention pooling layer. The OptiMol ligand encoder is a 3 layer GCN with 56 hidden dimensions. The decoder heads which take the concatenated RNA and ligand embeddings is a 3-layer multi-layer perceptron with 32 hidden dimensions at each internal layer. We use a dropout probability of 0.2 and apply batch normalization at each layer. Encoder models and data loaders are implemented in DGL [62] and training is done in PyTorch [40].

We train our networks using the Adam optimizer for 1000 epochs `Compat` setting, and for 20 epochs in the `Aff` setting.

E Fingerprint Prediction

To compare to RNAmigos, we use two ligands sets of decoys for our pockets. The *pdb* set is the set of 264 co-crystallized binders present in the PDB. The *decoyfinder*, consists in non-binders generated for each active using DecoyFinder[9]. For fairness, in this section, we follow the original validation protocol, except for removing ligands based on our filters. We use the *DecoyFinder* and PDB ligands sets as decoys sets and emphasize that at this step, we ignore the ChEMBL compounds. For each pocket in the test set, we perform a virtual screening and aim to retrieve the native binder among decoys. We present the results in Table 6.

	DecoyFinder	PDB Ligands
`RNAmigos`	0.609± 0.294	0.627± 0.284
`fp`	0.821± 0.272	0.856± 0.219
`fp + pre`	**0.858± 0.253**	**0.872± 0.210**

Table 6: AuROC values computed on *DecoyFinder* and PDB decoy sets. We compare those values for the original RNAmigos results and our models in the `fp` setting, with and without pretraining.

On this more challenging filtered data set, we reproduce the ability of `RNAmigos` to retrieve a binding affinity signal. Moreover, we see that the proposed enhancements yield a significant performance gain of over 25% AuROC points on the *DecoyFinder* decoy set, achieving a new state-of-the-art on this task. We observe that as expected, pretraining our RNA encoder boosts the performance of our method.

F Comparison between pocket and prediction similarity

We compare the similarity of our prediction for two pockets to the ones between these pockets assessed with RMscores. To compare our predictions, we compute a Jaccard score over the set of our top 5% predictions. We order pockets using a 1Dt-sne to find a smooth ordering and displayed this comparison in figure 10. We can see that there is a group of similar pockets, yet below our 0.7 cutoff, that also results in a group of similar predictions. However, our prediction are similar but not identical for related binding sites, indicating that it is able to be specific.

F.1 Model Ensembling

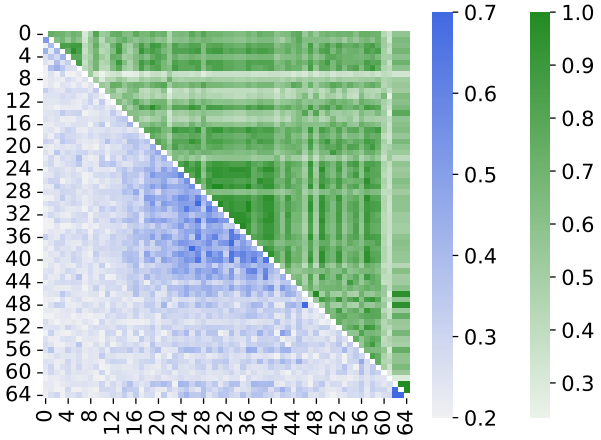


Figure 10: Comparison between RMscores of pockets (lower diagonal) and similarity of predictions (upper one).

	-	Compat	Aff	rDock
Compat	95.82 ± 0.90	96.51 ± 0.54	98.95 ± 0.41	**99.50 ± 0.15**
Aff	93.85 ± 0.40	-	93.99 ± 0.40	96.28 ± 0.18
rDock	95.93	-	-	95.93

Table 7: AuROC obtained by combining two methods