Semi-supervised learning and large-scale docking data accelerate RNA virtual screening

Juan G. Carvajal-Patiño[† 1,3], **Vincent Mallet**[† 2], **David Becerra** [1,3], **L. Fernando Niño V.** [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]D-BSSE, ETH Zürich, Switzerland
[5]MLSB, Max Planck Institute of Biochemistry, Germany
[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

RNA-small molecule interactions play a critical role in biological processes and have emerged as attractive targets for drug discovery. Despite recent enhancements in RNA molecular docking for virtual screening, the computational cost of docking remains prohibitive. Previous work has shown that using a proper graph representation of RNA structures and machine learning enables us to quickly get enrichment factors in small molecule ligand prediction. However, the performance of this approach is limited by the paucity of available structural data including interaction with a small molecular compound.

In this paper, we propose a competitive virtual screening pipeline to rapidly identify candidate ligands. We conduct large-scale docking experiments as a data augmentation and leverage unsupervised pre-training techniques for our RNA and ligand encoders to compensate for our reduced data. Our pipeline can also be used in combination with molecular docking, either by simply prioritizing compounds or by mixing the predicted score with the docking score.

We show that our new model widely outperforms previous ones. We also introduce a virtual screening test case for RNA and evaluate the performance of our tool in a realistic drug discovery setting. Our affinity prediction model shows native ligand recovery rates competitive with full docking experiments with runtimes in seconds. When the affinity prediction model is used to prioritize docking candidates, we show a 1200% acceleration of native ligand recovery with no performance loss. Finally, we show that using our model in combination with docking enhances virtual screening results with almost no overhead. 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.

1 Introduction

The development of small molecule therapeutics is a long-standing pillar of modern medicine [8]. Once equipped with a biological target, small-molecules can be experimentally tested for binding activity against this target. Due to the cost of binding assays, only a limited number of compounds can be tested out of chemical libraries with tens of millions of compounds - and an estimated chemical space of 10^{20} up to 10^{40} drug-like compounds [11]. Virtual screening (VS) techniques are used to select a promising shortlist of candidate compounds from chemical libraries. These techniques are classified into *ligand-based* and *target-based* methods. Ligand-based methods use similarity to known active compounds to find new candidates. On the other hand, receptor-based methods use the structure of the binding sites of a target, allowing the identification of binding modes different from those obtained with the known ligands [14]. Molecular docking, which explicitly models the target and ligand atom sets and measures various energy terms to search for the lowest energy binding pose, is a popular target-based technique to estimate the strength of binding of a compound to a binding site; however, it is also computationally expensive [22, 46]. Some recent approaches try to circumvent expensive energy calculations and use machine learning to leverage the abundant protein structural data available to identify patterns of likely binders. They appear as a promising alternative to docking for proteins [2, 36, 17, 31].

Most commercially available drugs target proteins, which represent 1.5% of the human genome and only a small fraction of it is druggable. By contrast, up to 70% of the human genome is used to produce non-coding RNAs [41]. RNA molecules play critical roles in a wide range of biological processes and are increasingly recognized as promising targets for the development of novel small molecule therapeutics [12]. However, identifying small molecules that can bind selectively and effectively to RNA is a challenging task because RNA molecules are more flexible than proteins and hence less RNA structural data is available. Despite this, interest in RNA drug discovery is growing rapidly [12, 16, 9, 1].

Computational methods for RNA drug discovery have advanced in recent years mainly along two lines: explicit RNA-ligand complex modeling and implicit/learned scoring. Along the first line are nucleic-acid-specific docking tools and scoring functions such as rDOCK [33], AnnapuRNA [37], and Autodock VINA [39] which use knowledge-based potentials to solve for the optimal pose given a ligand and target. These methods offer higher confidence screening results but come at the cost of high computational demands : time needed to evaluate a single pocket-ligand pair with a docking software is on the order of minutes. Attempts at accelerating this process use statistical or heuristic physics-based scoring approaches such as LigandRNA [29] and RLDOCK [38]. On the implicit side, methods try to further increase throughput by learning complex statistical associations between RNA and small molecules using machine learning methods. Here, tools such as InfoRNA [10] match a query secondary structure RNA with a library of structure motifs associated with known binders in an experimental assay. RNAmigos [24] made the first attempt at explicitly modeling the tertiary structure of the pocket to screen compound libraries and showed the potential of graph representations based on the Leontis-Westhof nomenclature of base pairs in 3D [21]. Ligand-based screening approaches are also taking shape, with ROBIN [43] which uses an in-vitro assay to screen large libraries of compounds against a set of RNA targets from which a ligand classifier tries to distinguish RNA binding from non-RNA binding compounds.

In this paper, we introduce the first machine learning based method for virtual screening on RNA that is competitive with molecular docking in a fraction of the time. Whilst our previous work provided interesting insights on RNA binding sites representation[24], the paucity of RNA small molecule binding data impeded the performance of this approach, hindering its use in real drug discovery pipelines. Here, we address this data paucity by generating docking values and 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. Finally, we introduce a realistic virtual screening validation and benchmark the performance of our tool against molecular docking.

Despite running in seconds instead of hours, the proposed approach retrieves similar enrichment factors (top 1.9% vs 1.2% of the candidate ligand list) to molecular docking. Moreover, using the tool and docking jointly, we are able to enhance docking virtual screening enrichments either by cutting computation costs by a factor twelve for the same enrichment or by improving enrichment from 1.2% to 0.8% without significant computation overhead.

2 Methods

2.1 Dataset Construction

The first step of our pipeline is to create a relevant dataset for training and evaluating our models. Training data has four main components: 1) a supervised dataset of binding site-native ligand pairs from the PDB, 2) synthetic affinity data from large-scale docking experiments, 3) local substructures sampled from a non-redundant set of RNA structures for unsupervised learning, and 4) a large library of chemical compounds. These datasets are summarized in Figure 1.

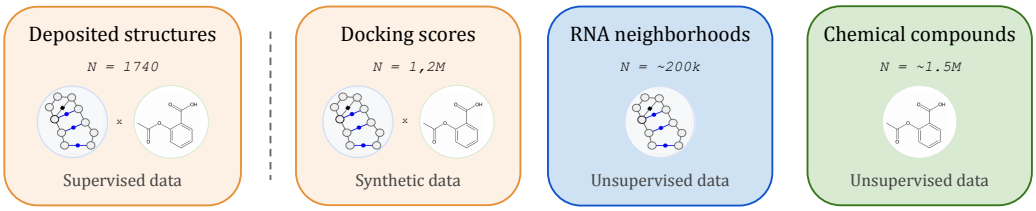


Figure 1: Datasets used in this paper. Beyond the supervised data available in the form of 1740 deposited structures of a ligand and its binding site, we use molecular docking as synthetic data and unlabeled datasets of 1.5M chemical compounds and 200k RNA residues neighborhoods.

Constructing a Dataset of Binding Sites

To construct our dataset of binding sites, we downloaded all RNA structures published in the RCSB Protein Data Bank (PDB) [32] and we filtered these structures into a set of RNA-small molecule complexes containing RNA and at least one ligand. We then extract RNA binding pockets 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. Pocket-ligand pairs (systems) are filtered based on their ligands, following Hariboss rules [26] (e.g. removing ions, crystallization artefacts, etc.). The final dataset consists of **1740 binding sites** associated with **264 unique ligands**.

We then used RNAGlib [23] to obtain a representation of the structure of our binding sites in the form of ABPN 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 [21]. This representation was shown to give a superior performance in [24], though using undirected graphs which loses backbone geometry. In this work, we build directed graphs as they are relevant to disambiguate some asymmetric non- canonical interactions.

Building a Library of Docking Scores

We build three ligands sets to screen against our pockets. The first set of small molecules is the set of 264 co-crystallized binders present in the PDB. For each of those compounds, we obtain non-binders using DecoyFinder [6] to simulate a virtual screening without docking (following RNAmigos), resulting in a second set. The last set of compounds is used to simulate an actual virtual screening. We used the ChEMBL [13] database to select ligands with physico-chemical characteristics plausible in a drug design setting. We picked a diverse set of ChEMBL compounds using the MaxMin algorithm (we used the RDKit implementation [19]) based on Tanimoto distance between fingerprints. We complement this procedure with 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* below five. We follow that procedure to retrieve a fixed set of 500 compounds. Then, for each binding site, we ensure that up to 100 compounds have a Tanimoto Coefficient greater than 0.6 with the native ligand by sampling additional ligands if necessary.

We assessed two candidate docking tools, rDock [33] and AutoDock Vina [39] which were identified in a recent review of nucleic-acid docking tools [22]. In addition, we used AnnapuRNA [37], a knowledge-based scoring function, as an independent referee to compare these candidate software in terms of native pose identification. We performed self-docking and cross-docking experiments with this software to compare their capabilities for RNA molecules and present the results in the

Appendix (section A.2.1). Eventually, we chose to use rDock to train our models as it showed the best performance.

We perform docking for two reasons: to use the docking results as a data augmentation, and to obtain virtual screening results. We thus dock the 264 native binders and the 500 ChemBL compounds in our 1740 binding sites, resulting in over 1.2 million values (runtime details are presented in A.2.2). This enriches our original scarce dataset of experimental binding sites. We additionally rely on all RNA residue neighborhoods and a curated dataset of ZINC[18] compounds to pretrain our RNA and molecule encoders, as will be detailed below. The whole of the data we will be using is presented in Figure 1.

2.2 Models

Equipped with this dataset, we are now ready to train models to complement molecular docking with an increased throughput, building over RNAmigos [24]. We use the ABPN graph representation, and make it more powerful by allowing the graph to have directed edges, resolving an ambiguity for asymmetric edges such as Sugar-Hoogsteen (SH). 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) [34] 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 $\phi_\mathcal{G} = f_\theta(\mathcal{G})$. We then propose three strategies to model binding sites and ligands interactions, depicted in Figure 2.

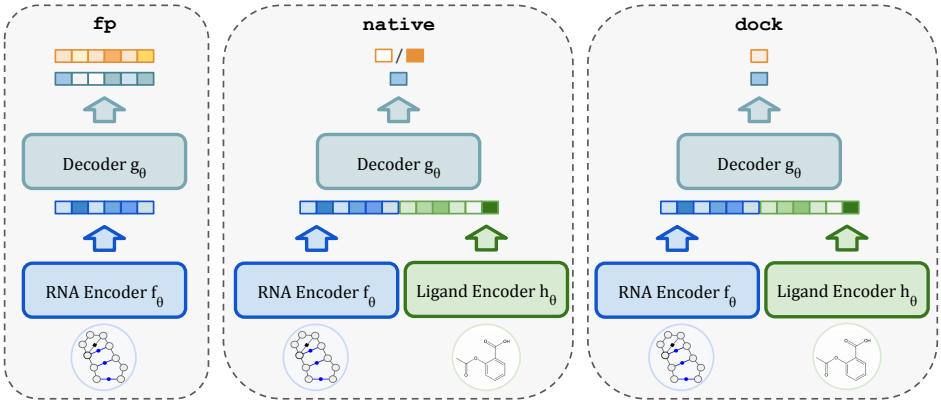


Figure 2: Three proposed models to learn RNA-ligands interaction. In the `fp` setting, we input the binding site to a graph encoder, and decode the resulting embedding into a predicted fingerprint for this binding site. In the other settings, we additionally encode candidate ligands structures and use our decoder to predict a compatibility score. For the `native` setting, the score is supervised with a binary classification into native or decoy, whereas in the `dock` setting, this score tries to approximate the docking score for the given binding site and ligand pair.

In our first setting (`fp`), this graph representation $\phi_\mathcal{G}$ is fed through a decoder g_θ to predict a fingerprint of dimension d, following RNAmigos[24], so that our model can be written as $M_1 = g_\theta \circ f_\theta$ with $M_1(\mathcal{G}) \in [0,1]^d$. We used a Multi-Layer Perceptron (MLP) for the decoder and train our networks using the norm two distance between our prediction and the fingerprint of the native ligand $\mathcal{F}$, so that the loss can be written as $\mathcal{L}_2(\mathcal{G}, \mathcal{F}) = \|M_1(\mathcal{G}) - \mathcal{F}\|^2$. One can then search a library of compounds for binders by finding the closest neighbors to the predicted fingerprint using Tanimoto similarity. We will refer to this setting as `fp`.

For our second setting (`native`), we additionally use the ligand $\mathcal{F}$ as input to our model. To represent the ligand, we use its molecular graph with the features introduced in OptiMol [5]. We then feed this molecular graph into the OptiMol ligand encoder, denoted as h_θ, to obtain a ligand embedding. 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 second model as $M_2(\mathcal{G}, \mathcal{F}) = g_\theta(f_\theta(\mathcal{G}) \bigoplus h_\theta(\mathcal{F})) \in \mathbb{R}$. We then train this network to output one for native ligands and zero for decoys, by minimizing a Binary Cross Entropy (BCE) loss, and refer to this setting as `native`.

In our third and last setting (`dock`), we keep the M_2 model described above, but we change the training procedure so that the interatomic term of the docking score becomes the prediction target. Splitting in this scenario is done such that the test set only contains unseen pocket-ligand pairs. We normalize docking scores to be between zero and one and train our network to approximate this value using an $\mathcal{L}_2$ loss. We refer to this setting as the `dock` setting.

Model pretraining In low data regimes which are common for tasks on RNA 3D structures, it has been shown that pre-training models on larger yet unlabeled datasets can benefit the performance on specialized downstream models [24]. We therefore pretrain our RNA encoder using the strategies proposed in [25] which are based on the metric learning framework [42]. The intuition behind such methods is to train the model to recognize geometrically similar structures according to a similarity function. It was shown to help convergence for downstream tasks where structure is relevant. Our directed graphs open the door to using more powerful similarity functions, and hence enhance pre-training.

Following this framework, we pretrain our ligand encoder using rooted subgraphs around all nodes of RNA non redundant data represented as graphs. This results in over 200k small RNA graphs for pretraining. Moreover, we use the pretrained ligand encoder proposed in Optimol [5]. This encoder was trained using a curated set of ~ 1.5M ChemBL compounds [30], using a generative model objective and taking molecular graphs as input. The specific model architecture and hyper-parameter settings are described in Section B.

2.3 Virtual screening

Virtual screening (VS) aims to sort a library of compounds such that the active ones are at the top of the list [20]. Typically, a VS is used to screen millions of compounds, as only a few thousands can be sent to wet-lab assays. As presented in Figure 3, to perform virtual screening with our tool, we simply compute the scores for all of our compounds and sort them based on this score. Our tool can be used in conjunction with docking, either by using our scores to quickly presort the compounds and prioritize promising ones or by blending our scores with docking scores.

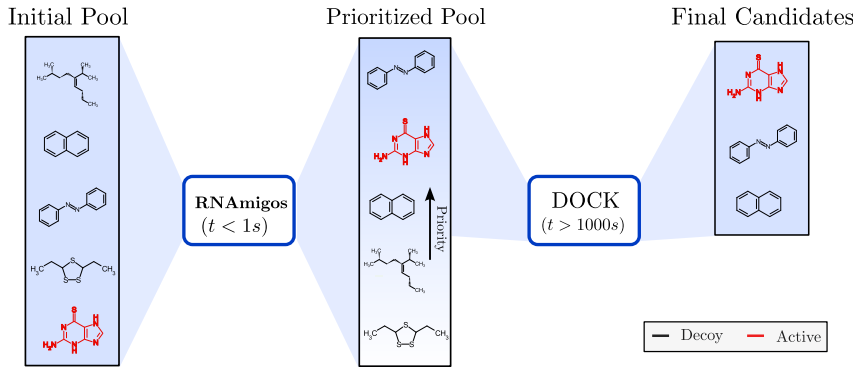


Figure 3: Illustration of the virtual screen procedure assisted by our model. First an initial unsorted pool of compounds is fed to our learned scoring models which quickly prioritizes this pool. Highest priority candidates are then the first ones passed to the computationally expensive docking software to identify the most promising compounds.

Metrics for VS quantify the gain in active compounds in a particular selection of ligands you obtain by employing a certain VS strategy compared to a random selection. The main VS metric are the well known AuROC and Enrichment Factors [7], where enrichment factors for a given VS method at a cutoff X gives the fraction of actives versus decoys (non-actives) in the top scoring $X\%$ relative to the overall active prevalence.

To our knowledge, no active/decoy set is available for RNA targets. We only have a single active for each target, namely the co-crystallized (native) ligand in the PDB. In this setting, given the rank R of the active ligand in a list of decoys of length N, the EF and AuROC simplify and only depend on the normalized rank R/N. Namely, the value of the EF at threshold x is either zero if $x < R/N$ or 1 otherwise, and the AuROC value is equal to $1 - R/N$ - since the ROC curve starts at zero and is

equal to one after this fraction. Therefore, in the following and for clarity, we will report AuROC scores and denote them as Mean Active Rank (MAR).

Because the number of drug-like compounds available in databases such as Enamine [15] is in the billions, slow running methods such as molecular docking are not tractable. Thus, we introduce a novel metric, denoted *efficiency*, to measure the expected performance if we were to stop our screening after a certain time budget is spent. We perform a virtual screening over a randomly ordered list of compounds and put processed compounds at the top of the list. We keep track of the MAR as a function of time and define the *efficiency* of a search as the area under the time vs. MAR curve. In this manner we can capture the trade-off between time expenditure and the quality of the search results in a single value. In this scenario, the performance of the tool heavily depends on the initial ordering of the list. For instance, on a list where the active compound is put last, any screening tool has a MAR of zero up until the very end. Hence, we repeat the computation over many shuffled initial orderings to get an expected efficiency.

3 Results

3.1 Proposed Dataset

An important result of this work is the generation of a diverse and updated data set containing binding sites, ligands, and docking scores. The complete data set of 1.2M molecular docking values can be downloaded at the following [2].

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 A.2.1.

We then turn to the analysis of our dataset of binding sites and ligands. In particular, we provide an analysis of the redundancy of our binding sites using RMScore[45] as a distance. We show that our pockets cannot be clustered and thus display a low redundancy (section A.2.3). We show that our ligands are diverse with a low pairwise Tanimoto score. Moreover, we show that the RNA ligands found in the PDB do not necessarily display drug-like features and often obtain worse docking scores than compounds derived from ChEMBL. This motivates our database and the integration of multiple data sources (section A.2.4).

3.2 Fingerprint Prediction

As a first validation, we want to compare the original RNAmigos paper to our fp setting that resembles the most theirs. The enhancements we propose are training on the updated data set, representing binding sites with directed graphs and updating the model architecture. 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 use the MAR metric introduced above as a metric and present the results in Table 1.

	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 1: Mean Active Rank (MAR) 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.

[2] https://github.com/jgcarvajalp/RNAmigos2

On this more challenging filtered data set, we reproduce that RNAmigos is able to retrieve a binding affinity signal. Moreover, we see that the enhancements yield a significant performance gain of over 22% MAR 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.

3.3 Virtual Screening Validation

We now turn to an evaluation of our results using our full dataset. For each binding site, we now use a decoy set containing ChEMBL compounds similar to the native ligand, a scenario closer to an actual drug design validation. On this new data, we train models in our three settings, and report the MAR metric value. We include rDock performance for comparison and present results in Figure 4. The evaluation set in this experiment is a held-out set of 128 pockets which do not appear anywhere in the training or validation data.

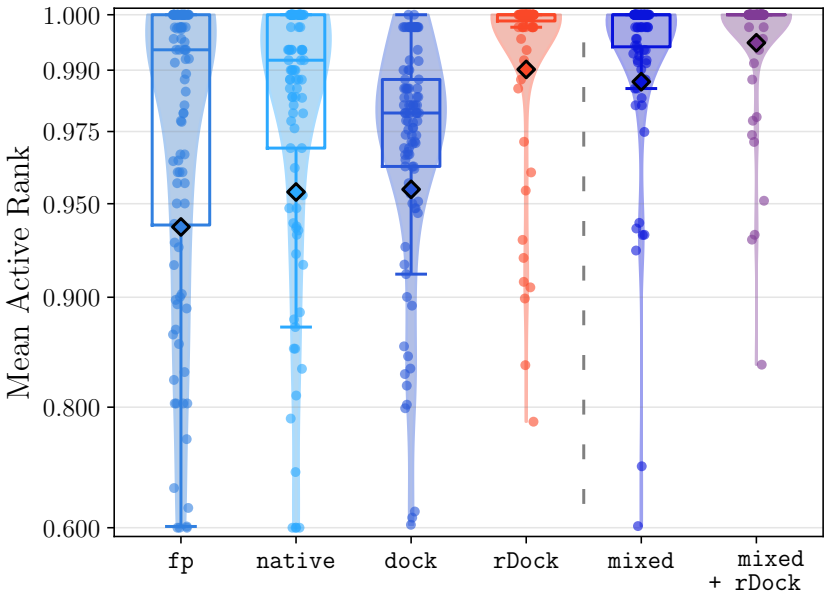


Figure 4: Distribution of the MAR performance of different virtual screening models, split by test pockets. Diamonds represent the mean performance. We compare models trained under our three settings with rDock (left of the dashed lines), an ensemble of these models and the average of this ensemble model and rDock.

We find that all the tested models have a mean MAR higher than 0.94. Note that such high values are common in VS, as one should remember that 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. Our results show that a strong VS signal can be extracted by our machine learning methods. Moreover, out of our three models, the dock model obtained the best results (with a value of 0.956) showing that docking scores are a useful oracle to train a machine learning model. Finally, rDock outperformed our models with a mean value of 0.99. 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 pocket.

3.4 Model ensembling

Since each of our setups is trained with its own modeling choices, by averaging their results, we expect to be able to mitigate their individual errors. Hence, we compute the enrichment factors obtained by averaging the outputs of our models and also possibly with docking. The weights of these averages are computed over the validation set. We have made all pairwise combinations and

report results in Table 2. We have also studied the combination of all our three models to create a model coined as *mixed*, that we also tried to combine with rDock and present results in Table 3.

	fp	native	dock	rDock
fp	0.940	0.970	0.983	**0.994**
native	-	0.954	0.979	0.992
dock	-	-	0.956	0.991
rDock	-	-	-	0.990

Table 2: Mean Active Rank (MAR) values obtained by merging two of our methods.

Method	MAR
mixed	**0.988**
mixed + rDock	**0.995**

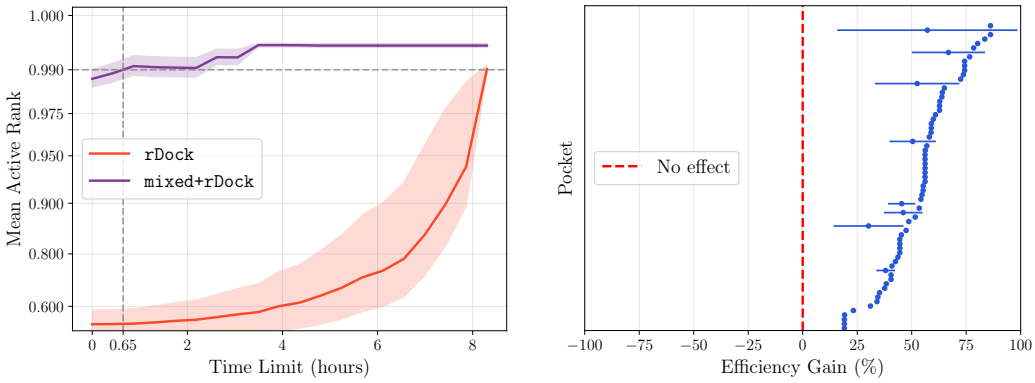
Table 3: Mean Active Rank (MAR) values obtained by merging our three methods (mixed) with and without rDock.

We see that combining models always leads to an increased performance, especially when mixing a model based on docking with a model based on native binding. Using our best mixing, we reach an impressive performance of top 1.2% rank for the mixed to be compared to a 1% performance for rDock. Moreover, we observe that when using the average of mixed and rDock, we reach a performance of top 0.5%, halving the error of rDock on its own.

3.5 Complementary Use of Docking and Learned Models

In the previous experiment, our best performing model needed about thirty seconds to obtain an enrichment factor of 0.988. The rDock models needed about 8 hours to run per pocket, achieving 0.990 when unmixed and 0.995 when mixed. Hence, our model is over three orders of magnitude faster than its docking counterpart, and achieves a close, albeit lower enrichment factor.

We put ourselves in a scenario where we have a limited time budget and want to assess the *efficiency* of our method, as described in section 2.3. We propose to compare the usual docking setting with one where compounds are pre-sorted according to our mixed score, and then sorted with the average of this score and the docking score (*mixed+rDock* setting). We expect our presort to help the docking tool review the promising compounds first and thus spend its time budget more efficiently. Results are shown in Figure 5a. Figure 5b reports the *efficiency* metric split by pockets.



(a) Timed Enrichment Factor experiment. We compare the average enrichment factor over time of our models.

(b) Gain in efficiency obtained by following the mixed strategy, split by pockets.

Figure 5: Time-budgeted virtual screen

In this novel approach the mixed model is able to reach the same MAR values as rDock in approximately 1.6 hours, more than twelve times faster than rDock. Moreover, we show that this performance boost can be expected for most of our pockets, and more importantly that this process never results in a waste of time (all pockets are to the right of the red vertical in Fig 5b). This experiment shows that using our learned models in conjunction with a docking tool is a promising avenue to accelerate RNA virtual screening throughput and hence open the door to the screening of larger libraries, potentially those coming from ligand generative models [4].

4 Conclusion

We address the timely question of *in-silico* RNA drug design. In contrast to its protein counterpart, the reduced amount of structural data limits the direct applicability of machine learning based approaches. Dedicated methods are needed as the RNA drug design field becomes more established.

In this work, we built a dataset of RNA binding sites structures, along with their ligands, and a curated set of drug-like compounds. To make up for the scarce data, we use RNA docking as a data augmentation strategy and rely on unsupervised pre-training techniques. We start by showing that this method outperforms existing machine learning models by a margin of over 22%. Then, we compare different models with docking and show that we obtain comparable enrichment factors, in a fraction of docking runtime. When scoring compounds with an average of docking scores and our models, we are able to halve the docking error. Finally, we propose a strategy to get the best of both worlds by using our tool as a preprocessing step for docking and show that this strategy reduces the computational costs by a factor twelve at no performance cost on average, and with a positive impact on all binding sites in our test set.

This paper is a step towards establishing machine learning on the structure of RNAs, and unlocking its power for RNA drug design. By publicly releasing all our datasets, source code and model weights, we hope others will propose more efficient methods and spark a community effort in this important direction. Moreover, emerging techniques such as generative models of ligands are drastically increasing the size of the libraries of potential drugs. This is a throughput that the staggering computational needs of docking tools cannot match. We believe our tool could be used as a high throughput, high recall, filtering step for virtual screening in the large library settings.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C. M. Connelly, M. H. Moon, and J. S. Schneekloth. The emerging role of rna as a therapeutic target for small molecules. *Cell Chemical Biology*, 23:1077–1090, 2016.

[9] 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.

[10] 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.

[11] Peter Ertl. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *Journal of chemical information and computer sciences*, 43(2):374–380, 2003.

[12] 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.

[13] 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.

[14] Aleix Gimeno, María José Ojeda-Montes, Sarah Tomás-Hernández, Adrià Cereto-Massagué, Raúl Beltrán-Debón, Miquel Mulero, Gerard Pujadas, and Santiago Garcia-Vallvé. The light and dark sides of virtual screening: What is there to know? *International Journal of Molecular Sciences*, 20, 2019.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] Yuanzhe Zhou, Yangwei Jiang, and Shi-Jie Chen. Rna–ligand molecular docking: Advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(3):e1571, 2022.