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], **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 - Grupo de Investigación LISI - Carrera 45 # 26-85, Bogotá, Código Postal: 11001 - 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 of 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 2.8% on average on a structurally distinct drug-like test set. Those predictions are sensitive, but robust to pockets alterations, opening the door to its use on binding site detection methods outputs. Our model results in a ten 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 a mean enrichment factor at 1% of 2.93 on four unseen RNA riboswitches. To our knowledge, 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 [73]. For instance, about 2000 genes code for micro-RNAs, which in turn affect the expression of 60% of genes [18]. Despite this ubiquity, the first RNA-targeting drug, *risdipalm*, was only recently approved by the FDA [42], 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 [67], 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 [75]. In this context, RNA is increasingly recognized a promising family of targets for the development of novel small molecule therapeutics [16, 21, 13, 2], highlighting the need for efficient tools for RNA drug discovery [16].

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 [4]. Recently, many binding affinity prediction methods based on machine learning [5, 60, 23, 53, 70, 74, 37] 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 [39]. 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 [54] contains hundreds of thousands of experimental protein structures, and now hundreds of millions predicted structures from AlphaFold [27], only a couple of thousands are available for RNAs. Similarly, databases of protein-ligand interactions such as PDBBind [72] 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 [66].

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 [55], DOCK 6 [3], and RLDOCK [63]. To enhance the docking procedure, alternative pose scoring functions have been proposed such as AnnapuRNA [61], LigandRNA [51], and very recently FingeRNAt [65], and RLaffinity [64]. 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. Efforts have been made in accelerating docking. First, the ZINC20 [25] introduced a large pre-processed ligand database, allowing to skip some ligand preparation steps. In addition, some methods were proposed to deploy computations on a GPU, and manage to screen compounds in approximately 0.1s per compound, per GPU, without sacrificing accuracy [56, 78]. Finally, some protocols tune docking approximation parameters for a specific pocket on a subset of ligands, enabling a faster, less accurate screening running in approximately 0.8s per compound per core.

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 [14] and InfoRNA2.0 [15] 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 [44] 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 [34], 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 [44], 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 2.8% vs 4.1% 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 four-fold while improving the predicted rank of actives from 4.1% to 1.0%. This strong performance is found to be maintained over a structurally diverse and unseen set of targets and compounds, and upon pocket perturbation. Finally, we achieve strong enrichment factors and diverse retrieval of diverse leads 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

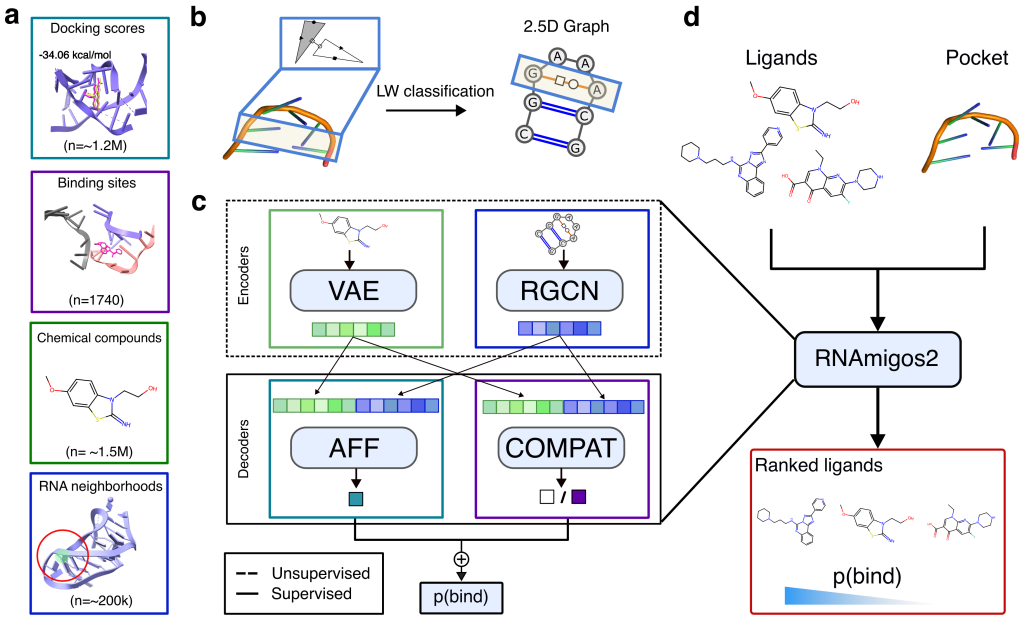

Figure 1: **a**: Datasets used in this paper. Beyond the supervised data available in the form of 1740 deposited structures of a ligand and its binding site, we use 1.2M molecular docking values as synthetic data and unlabeled datasets of 1.5M chemical compounds and 200k RNA residues neighborhoods. **b**: Representation of RNA structure as a 2.5D graph. Each binding site is encoded using Leontis-Westhof classification [34] 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 [35]. This representation allows us to capture key features of the RNA 3D architecture with a discrete mathematical object [34, 33] that suits well machine learning frameworks and was shown to be a useful biological prior for RNA cheminformatics applications [44]. The RNA encoder takes the 2.5D graph as input and learns to generate an RNA representation using self-supervised training schemes [45] 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 [9] and trained on large dataset of chemical compounds [24].

To train decoders, we extract 1,740 RNA-ligand complexes from the PDB[7] and group them in 436 clusters of similar binding sites we identified using RMAlign [80] with a similarity threshold of 0.75. *To our knowledge, this is the first rigorous structure-based split for RNA drug-target association prediction.* 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 [44] 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 A.2.

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 [19] 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 determined by the MinMaxPicker algorithm ([30]), 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 [59]. 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[80] 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. Native ligands of binding sites belonging to the same group are considered as positives for this group. Then, we performed a data splitting on the groups, ensuring no structural data leakage exists between our training set of 367 groups 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 0.75) 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 0.844 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 less than five seconds on a single machine, whereas the docking results were obtained after $\simeq$ 8 CPU hours per binding site.

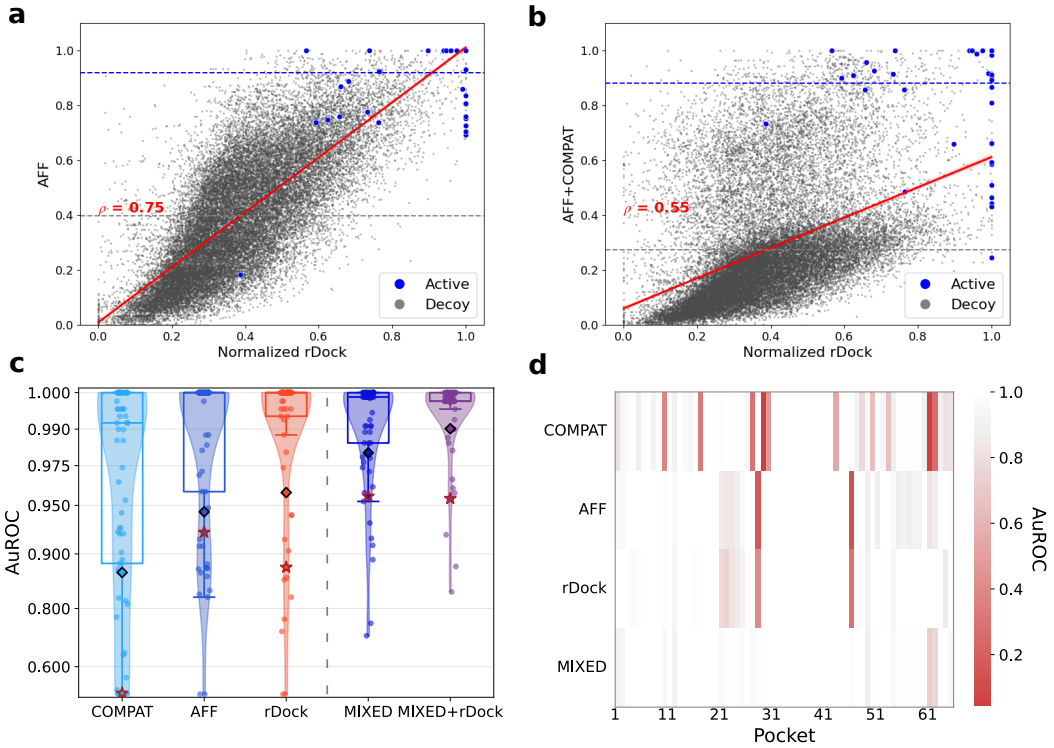

Figure 2: **a,b** Relationship between `Aff` (a) and `Mixed` predictions against rDock scores for decoys and native ligands in the test set. Horizontal lines indicate the mean of predicted scores for native and decoy ligands. **c** Distribution of the AuROC performance of different virtual screening models over test binding sites. Diamonds represent the mean performance and stars represent performance with a shuffled target. 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.

Interestingly, we find that the errors in VS are aligned for each binding site for our different models (see Figure 2**d**). While as expected, rDock and `Aff` errors align, the `Compat` trained on an independent data source has complementary error modes. This complementarity might originate from force field priors being too strong to accurately fit the data. This suggests a simple strategy for mitigating their individual errors, and we propose to ensemble each model's results by taking the max of their ranks to create a model coined as `Mixed`. Our averaging is reminiscent of Bayesian fitting, with a likelihood term and a physical prior term.

The `Mixed` scores now have a lower correlation of 0.55 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 an ensemble of two models trained on the same data source, and report AuROCs over 3 independent seeds for all pairwise combinations in supplementary Table 1. We reach a performance of 0.972 AuROC for the `Mixed` to be compared to 0.897 and 0.942 for the ensembled versions of our single modality models.

This performance gap proves the synergistic effect of our models. It also enables us to significantly outperform rDock (0.959) (Mann-Whitney U test p-value of 0.026). 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 machine-learning based RNA virtual screening tool.

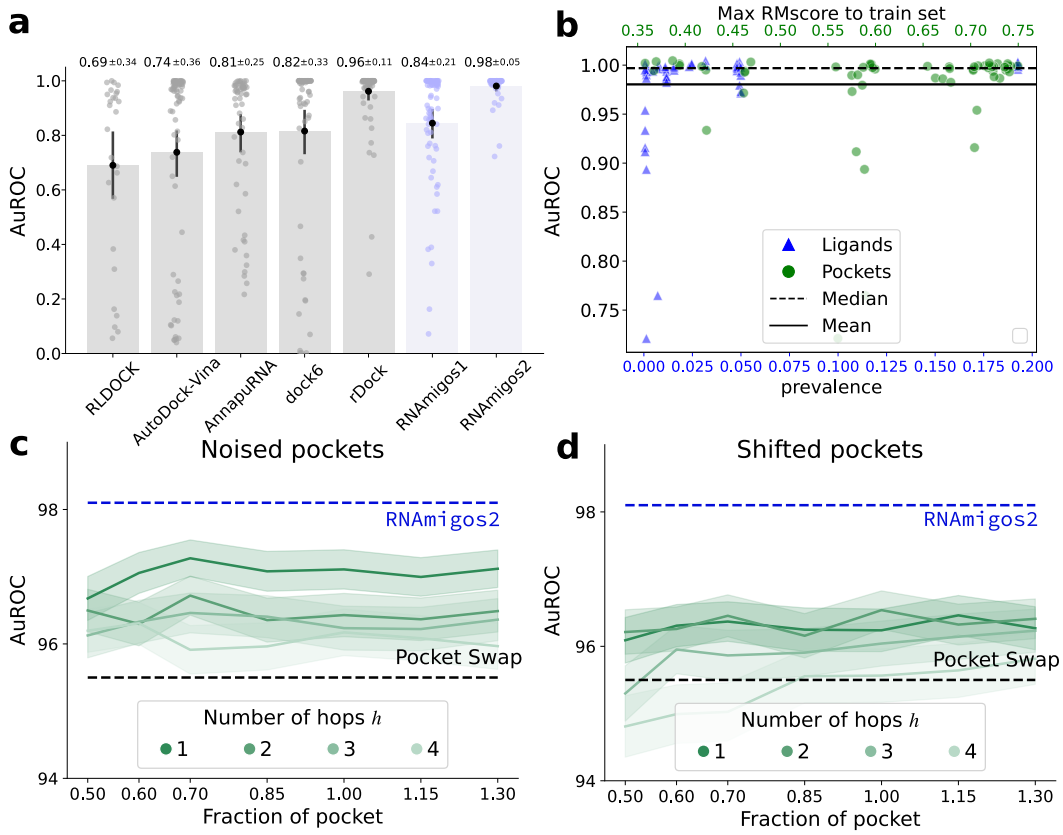

Figure 3: **a.** Virtual screening benchmark of RNAmigos1 and 2 vs state of the art docking tools on ChEMBL compounds against state of the art docking tools. Note, RLDOCK produced errors on 42 binding sites. We only report its performance for 27 binding sites, which we denote with RLDOCK*. **b.** Depiction 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. **c,d.** Mean and standard error of the AuROC for different perturbations. For each pocket p, we first expand it with h hops to get p_h, and either sample nodes in p_h (*noised*), or around a seed node at the boundary of p_h (*shifted*). The number of nodes we sample is computed as a fraction of the size of p. The sampling is replicated ten times. Unperturbed performance is shown in blue.

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[63], AutoDock+Vina[68], Annapurna [61], DOCK 6, [31] rDock [55] and RNAmigos1[44] on a virtual screen against ChEMBL compounds. As presented in Figure 3**a**, our method outperforms all methods with a wide margin 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 1. 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 [80]) 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.

Finally, we assess our model specificity and robustness. Specificity is known to be hard for drug-target interaction prediction. Indeed, decoy compounds tend to have a chemical bias, such that active compounds can be found by a simple QSAR model. An interesting swapping experiment is proposed in Volkov et al. [69]. It consists in shuffling target-ligands pairings, so that compound affinities for a target are associated with another target chosen at random. The authors manage to retain high enrichments for models trained of shuffled pairs, disproving their specificity. While some enrichment is expected based on general properties of active chemical compounds, a high specificity model should display as large a performance gap as possible on shuffled targets. We visualize the results of this experiment on Figure 2; by denoting the swapped performance with a red star. It can be seen that our `Aff` model has a small gap of 2.0 AuROC points, while rDock has a gap of 7.8 points. Thanks to a margin loss term correcting for this bias, we report a high pocket specificity in the `Compat` setting, where swapping induces a 41.7 point performance drop. This allows our `RNAmigos2` model to show target specificity, where swapping results in a drop from 98.1 to 95.6 AuROC.

Another key aspect of virtual screening is its robustness to the definition of the pocket. Indeed, current state-of-the-art binding site detection methods [46] retrieve 70% of the correct residues, and potentially miss the exact localization of the pocket. We conduct an experiment to assess the robustness of our tool with regard to imperfect pockets. We implement algorithms for sampling perturbed binding sites, precisely described in section E.2. For each pocket p, we first expand it with a breadth-first search of h hops around the original binding site, yielding the expanded pocket p_h. Then, either we sample nodes in p_h, yielding a wrong sampling in the right average localization (*noised*), or around a seed node at the boundary of p_h (*shifted*), mimicking errors in the binding site localization. The number of sampled nodes is computed as a fraction of the size of p, ranging from 0.5 to 1.3, and the sampling is replicated ten times. We present results in Figure 3**c,d**.

Our methods display some sensitivity to those perturbations, in line with the swapping experiments results. However, when sampling nodes in the close vicinity of the pocket (*noised*, h=1), we are able to retain most performance with about 0.7% AuROC drop. When growing h, the pocket becomes more scattered, smoothly decreasing our results to reach a 2% drop. When perturbing the localization, the drop ranges from 1.8 to 3%, which is more significant, but remains a strong signal considering that the most disrupted pockets barely overlap with the original pocket. At the highest level of perturbation, the performance can be worse than one obtained over swapped pockets, which can be explained by perturbed pockets that do not correspond to a pocket anymore. Finally, we observe a remarkably stable behavior with regards to the fraction of the pocket f, which is a very promising result for synergic usage with binding site detection models. To explain this stability, remember that we expand our binding sites with context before embedding it. Hence, we believe that our model implicitly learns to detect the exact binding site from an approximate location, explaining its robustness.

`RNAmigos2` predictions consistently boost virtual screening efficiency

Our model outperforms docking accuracy in a fraction of the time. However, since rDock displays a strong performance, we also investigate mixing the `RNAmigos2` model with docking, and show it enhances performance from 98.1% to 99.0%, halving the error of these models. Interestingly, this also increases the pocket specificity of our model. We refer to this more accurate, but more expansive strategy, as `RNAmigos++`, and plot its performance in Figure 2**c**.

In practice, we have a limited time budget. revExisting alternatives can bring this run time to 0.1 gpu seconds/compound, but rely on the availability of a GPU, or to 0.8s core second/compound at the cost of a less accurate result. We found docking with rDock to take approximately 65 core seconds/compound. Hence, we cannot afford to scan large scale databases with docking methods, which introduces a trade-off in virtual screening between quality of the scoring and the amount of computational time expended. In contrast, `RNAmigos2` needs 5.8 core ms/compound on raw smiles, and 1.6 core ms/compound on precomputed ligand graphs (a one-time overhead cost), which can be useful for projects involving several targets. This scoring can be considered negligible compared to docking run time. Hence, in a drug design setting and under a compute budget of one core day, conventional docking enables the screening of about $\sim$ 1400 compounds while `RNAmigos2` enables screening $\sim$ **15.7M 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: we report VS performance of the different methods for different time budgets (x-axis). For a time budget lower than the maximum, only the first compounds are sorted using our methods, whilst remaining ones are kept in their initial order at the end of the list. Introducing the *efficiency* of a VS method as the area between the curves of AuROC through time for this method and rDock (further developed in Methods), the *efficiency* of our hybrid method is shown in Figure 4C.

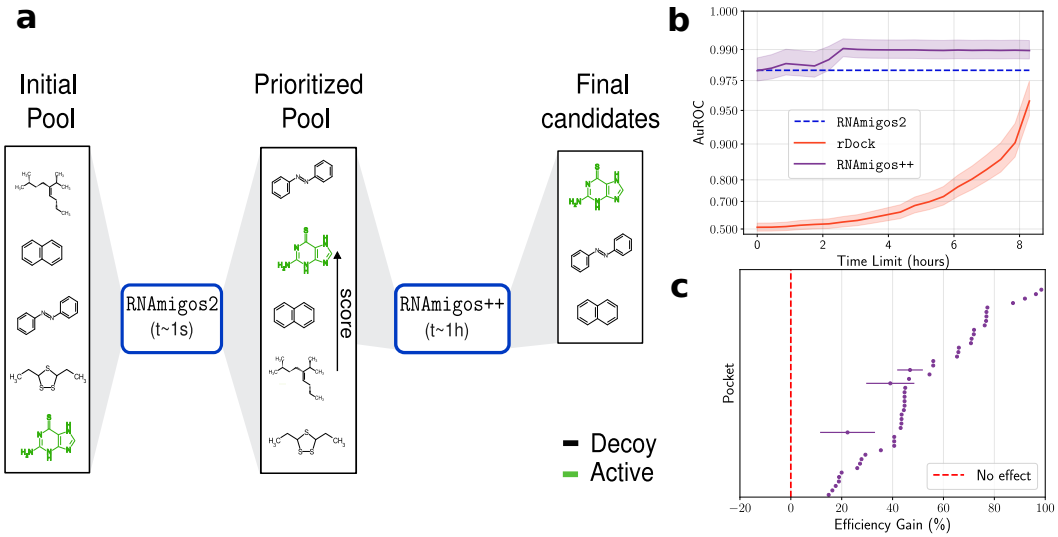

Figure 4: **a.** Illustration of our hybrid strategy. First, we sort the compounds using RNAmigos2 with a high throughput. Then, rDock is run only on the best scoring compounds and used to replace our docking surrogate score. **b.** AuROC of different models, averaged over binding sites, as a function of time budget spent. AuROC results were run with ten shuffled replicates of the initial order of the list. Note the logarithmic scale on the y-axis, that gives the impression of acceleration for rDock screening results. It is actually a linear interpolation between random and final performance. **c.** Efficiency metric split by binding sites, each line corresponds to a pocket.

This benchmark highlights that our methods obtain excellent results much faster than with a docking-only approach. Using RNAmigos2 in isolation, we get almost the same AuROC as rDock in 11,000 less time. Screening all compounds with both tools cuts rDock error by a factor four. Crucially, the boost of performance with RNAmigos++ over the RNAmigos2 model is achieved after screening only one fourth of the compounds. Finally, we show that we improve the performance for almost all binding sites, which indicates that our protocol never results in a waste of time. These results suggest that our method could allow the screening of larger libraries, potentially those coming from ligand generative models [8].

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 36 nucleic acid targets, known as ROBIN [77]. As in preceding experiments, we seek to identify the active compounds for each target which in this case are the range of 163 to 197. To make predictions using RNAmigos2, we identify all RNA targets in the ROBIN database which have a perfect sequence identity to an RNA chain in the PDB and pass the corresponding 3D structures as input to the model. Importantly, we ensure that all RNAs collected were not present in our training or validation sets, making this a blind testing scenario. As RNAmigos2 requires a binding site as input, we use the native ligand in the PDB as before to define

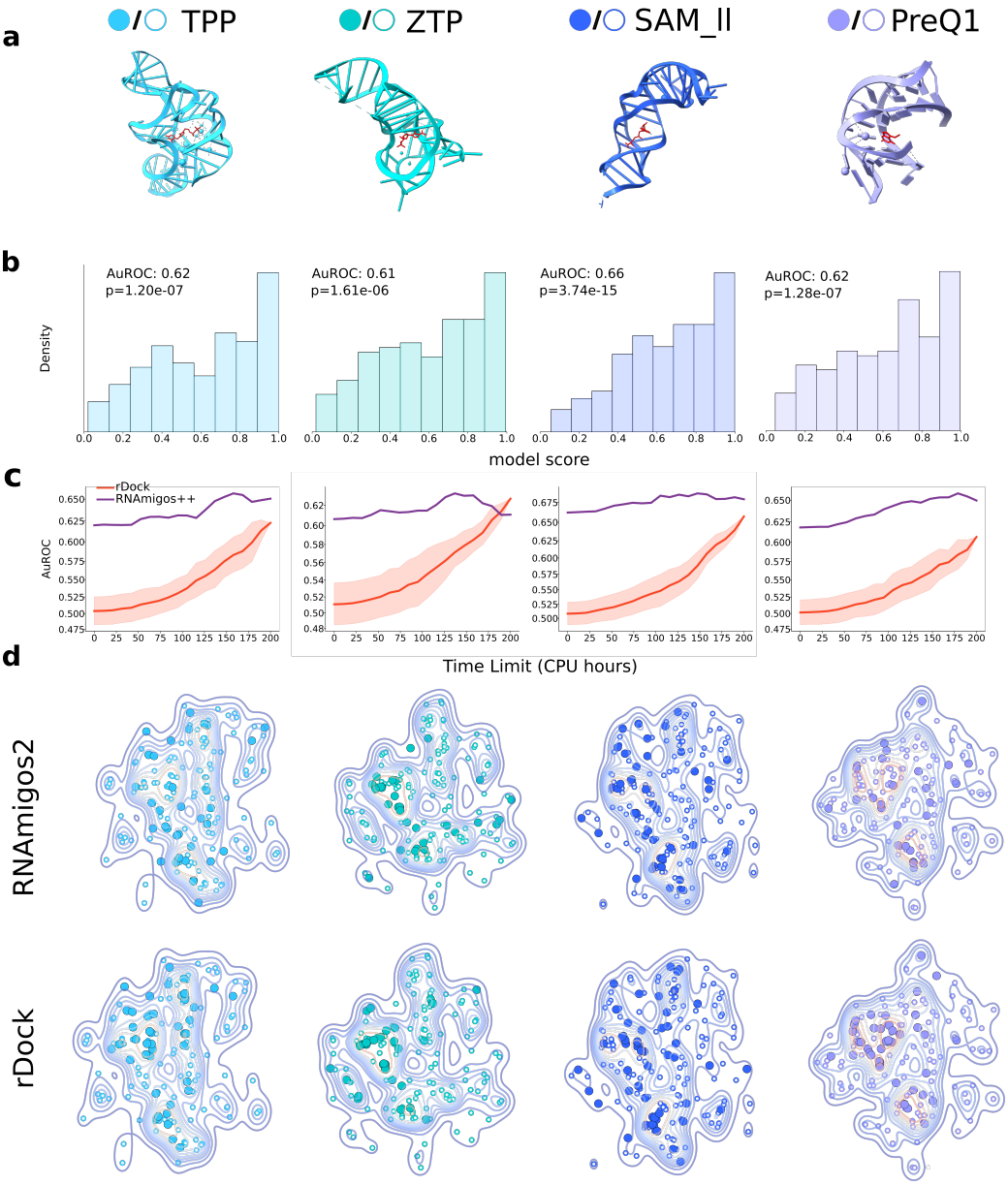

Figure 5: **a.** Visualization of the structure of the riboswitches targets in the ROBIN dataset used in this experiment. **b.** `RNAmigos2` ranking of active compounds in the ROBIN database for each target. P-values reported are from the Mann-Whitney U test for comparing distributions of ranks of active and inactive compounds. **c.** Time-limited benchmark on each ROBIN target using `RNAmigos2` to prioritize compounds for rDock. Scheme follows previous Figure **d.** Chemical compounds selected (best 10%) by different methods, shown in the tSNE space of ROBIN active compounds. Large, filled circles correspond to selected actives and hollowed ones are non-selected actives. A kernel density estimate is shown beneath the points.

the input region. This results in four binding sites: the TPP , ZTP, SAM_ll and PreQ1 riboswitches, (PDB ids: `2gdi`, `5btp`, `2qwy` and `3fu2`), visualized in Fig.5a.

We computed the `RNAmigos2` scores of all 24,572 assayed molecules for all pockets in about two CPU minutes. Our first finding is that `RNAmigos2` yields positive enrichments of active compounds in all four targets (AuROC reaching 0.66 and enrichment factors up to 5.09), significantly distinguishing decoys from actives (Fig 5b, Mann-Whitney U Test scores range 3.73×10^{-15} - 1.61×10^{-6}) and

despite both the compounds and targets being unseen during training. Remarkably, performance of RNAmigos2 on its own is on-par with rDock while running in a fraction of time (2 core minutes instead of 1000 core hours ; see Fig. 5, and Supplementary Table 3 for full results). Combining our predictions with rDock (RNAmigos2 ++), though more time-consuming, achieves higher accuracies than either tool on its own in 3 of 4 pockets, and always improves screening efficiency (Fig. 5c).

Beyond AuROC and enrichment factors, we report that *using RNAmigos2 augments the chemical diversity of retrieved actives*. Retrieving diverse actives is a key feature of virtual screening methods, providing more opportunities for hit optimization as well as lowering the risk of systematic failure on later tests [81]. Thus, two actives from different chemical scaffolds can be more valuable than five related active compounds, which is not measured by AuROC. In Supplementary Figure 12, we show actives retrieved by different methods. It is evident that our Aff model focuses on main clusters, which represent generic RNA binders and allows this model to get high AuROC scores. This also holds partly true of rDock, but is in contrast with our Compat model that seems to focus on on binders specific to a pocket, in line with its margin loss training. This equips RNAmigos2 with enhanced diversity compared to rDock, as measured by mean Tanimoto distance to the closest neighbor, as well as by the Hopkins statistic [22]. More importantly, we find in Supplementary Fig 7 that the distribution of selected actives by RNAmigos2 and rDock show notable Earth Mover's Distance [50, 17] shifts of (around 0.3) suggesting we explore different areas of chemical space. This is illustrated on ZTP for instance, where the left cluster is only found by rDock but the bottom right part is retrieved only by our tool. This is an opportunity for RNAmigos++ that displays the highest diversity *on all targets*. We believe having several fast scoring functions with complementary sensitivity is an exciting avenue for obtaining truly diverse and high-accuracy virtual screening protocols.

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 model runs more than ten thousand times faster than docking, it replicates its performance for diverse targets. This result is sensitive to pocket identity, but robust enough to pocket perturbation to be used in conjunction with modern pocket mining algorithms [62, 46]. Moreover, by combining our model with actual docking scores for the top-scoring compounds, we manage to cut docking error by four, in one fourth of the time. Finally, we establish our tool's performance on an independent large-scale (25k compounds) *in vitro* binding screen, and show that it provides 2.9 enrichment factors in two CPU core 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, as well as for the modeling of binding site flexibility [41]. An interesting direction for future work is to investigate the use of other docking tools to train other model, resulting in fast surrogates with potentially different error modes resulting from different scoring functions. We envision tools such as RNAmigos2 will play a synergistic role with rapidly emerging RNA-centric technologies for molecular design [26] 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 [43].

Data & Code Availability

All source code needed to fully reproduce results, model weights, and training data are available on `https://github.com/cgoliver/rnamigos2`. We also maintain a Colab notebook for convenient access to model inference. Following peer-review datasets and weights will be uploaded to Zenodo.

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] William J. Allen, Trent E. Balius, Sudipto Mukherjee, Scott R. Brozell, Demetri T. Moustakas, P. Therese Lang, David A. Case, Irwin D. Kuntz, and Robert C. Rizzo. Dock 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, 36:1132–1156, 6 2015.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

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

[19] 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.

[20] 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.

[21] 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.

[22] Brian Hopkins and John Gordon Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

[23] 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.

[24] 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.

[25] John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073, 2020.

[26] 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.

[27] 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.

[28] 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.

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[30] 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.

[31] P. Therese Lang, Scott R. Brozell, Sudipto Mukherjee, Eric F. Pettersen, Elaine C. Meng, Veena Thomas, Robert C. Rizzo, David A. Case, Thomas L. James, and Irwin D. Kuntz. Dock 6: Combining techniques to model rna-small molecule complexes. *RNA*, 15:1219–1230, 6 2009.

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

[33] 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.

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

[43] 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.

[44] 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.

[45] 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.

[46] FP Panei, P Gkeka, and M Bonomi. Identifying small-molecules binding sites in rna conformational ensembles with shaman. *Nature Communications*, 15(1):5725, 2024.

[47] 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.

[48] 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.

[49] 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.

[50] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Diogo Santos-Martins, Leonardo Solis-Vasquez, Andreas F Tillack, Michel F Sanner, Andreas Koch, and Stefano Forli. Accelerating autodock4 with gpus and gradient-based local search. *Journal of chemical theory and computation*, 17(2):1060–1073, 2021.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-dock: Gpu-accelerated docking enables ultralarge virtual screening. *Journal of chemical theory and computation*, 19(11):3336–3345, 2023.

[79] 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.

[80] 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.

[81] Tian Zhu, Shuyi Cao, Pin-Chih Su, Ram Patel, Darshan Shah, Heta B Chokshi, Richard Szukala, Michael E Johnson, and Kirk E Hevener. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis: Miniperspective. *Journal of medicinal chemistry*, 56(17):6560–6572, 2013.

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)[54] 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. Additionally, in the same manner as [44], we discard binding sites where more than 40% of the total residues in the 10Å zone of the ligand come from protein chains. We additionally remove complexes with ligands that did not follow rules introduced in Hariboss[47] (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[40] to obtain a representation of the 3D 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. Each node also contains a 640-dimensional feature vector from the RNA-FM language model [12] which was trained on a large corpus of RNA sequences and provides a strong evolutionary signal. Edges represent interactions between nucleotides, split in interaction types that follow the Leontis–Westhof classification[35]. Briefly, the Leontis-Westhof classification is a 12-way discrete categorization of base pairing interactions according to orientation of interacting bases [35]. This classification is known to capture salient geometric characteristics of the RNA, and thanks to its graph-based nature, it is highly amenable to geometric deep learning frameworks [35, 45]. An undirected 2.5D graph representation was shown to give better performance[44] than traditional representations for machine learning. In this work, we additionally use directed edges, based on base-pair orientation, 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 [36], 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[24] compounds, introduced in [52] 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[19]. 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 [59]. We complement this procedure with the MaxMin algorithm based on Tanimoto distance between fingerprints, using RDKit[30], 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 three candidate tools, rDock [55], DOCK6 [31] and AutoDock Vina[68] which were identified in a recent review of nucleic-acid docking tools[38]. rDock was developed for NA-ligand docking, DOCK 6 is an extended version of DOCK 5 specifically optimized for RNA and AutoDock Vina was used without any specific optimization for NAs. In addition, we used AnnapuRNA[61], 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 D. 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.

Data splitting

We compute the RMscore[80] of all pairs of binding sites in our dataset, and report their distribution in supplementary Figure 10a. 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 10b, 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. We use a neural network encoder f_θ based on Relational Graph Convolutional Network (rGCN) [57] layers. Given the set of all relations $\mathcal{R}$, we denote as $\mathcal{N}_r(i)$ the set of neighbors of node i connected by relation r. The parameters of this layer l are $r+1$ matrices $W_r^{(l)}$ with the convention that self loops have an index 0. The node representation of node i at layer l, $h_i^{(l)}$ evolve to the next layer following the equation:

$$h_i^{(l+1)} = W_0^{(l)} h_i^{(l)} + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{c_{ir}} W_r^{(l)} h_j^{(l)}.$$

where $c_{ir} = |\mathcal{N}_r(i)|$. This update rule aggregates the representations of neighboring nodes based on their relational types and updates the target node's representation accordingly. We apply several rGCN layers to our input, resulting in node embeddings of dimension $d \in \mathbb{N}$ and a pooling step reduces those node embeddings to a single graph-level vector representation $\phi_{\mathcal{G}} = f_\theta(\mathcal{G})$.

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

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

in $\phi_{\mathcal{F}} = g_\theta(\mathcal{F})$. We use the architecture proposed in OptiMol[9], which is a stack of GCN layers (obtained by setting $r = 1$ in the rGCN equation above).

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 $M(\mathcal{G}, \mathcal{F}) = h_\theta(f_\theta(\mathcal{G}) \bigoplus g_\theta(\mathcal{F})) \in \mathbb{R}$. The specific model architecture and hyper-parameter settings are described in supplementary Section E.1.

Pretraining and training

Our encoders are pre-trained to serve as a strong initialization for the representation of the input binding sites and the small molecules. Our RNA encoder pretraining relies on the strategies proposed in[45] which are based on the metric learning framework[76]. Namely, given a similarity metric k, the network is trained to minimize

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

Using directed graphs open the door to using more powerful similarity functions k than the ones used in RNAmigos1, and hence enhance pre-training. In this work, the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are defined as subgraphs of whole RNAs, expanded from two root nodes. Then, we compute graphlets decomposition of these neighborhoods, and compute a pairwise distance matrix between graphlets of $\mathcal{G}_1$ and $\mathcal{G}_2$, using a graph edit distance adapted for RNA similarity. Finally, the k function uses the Hungarian algorithm[29] to match the nodes following this pairwise distance matrix, and returns the score of the match.

Our ligand encoder, g_θ, is pretrained following OptiMol[9]: in addition to its graph encoding, we encode molecular compounds as a SELFIES [28] string. The encoder is then trained in conjunction with a decoder, that aims to reconstruct this string from $\phi_{\mathcal{F}}$ following a Variational Auto-Encoder (VAE) framework. This decoder is not used in our work, but we used the weights of this encoder as the initial weights of our ligand encoder.

Our first decoder (`Compat`) uses this architecture and is trained to output 1 for native ligands and 0 for decoys, by minimizing a Binary Cross Entropy (BCE) loss over experimental complexes. This problem is very unbalanced because we have many more negative pairs than positive pairs. Hence, at each epoch, we sample one active and one inactive of each group and train only on those two examples per group. In addition to this loss, and in order to increase the specificity of our model, we added a margin loss $\mathcal{L}_m$ that aims to set apart the score of a ligand $\mathcal{F}$ with its active $\mathcal{G}_{\mathcal{F}}$ from the one of $\mathcal{F}$ with another pocket for which it is inactive $\tilde{\mathcal{G}}_{\mathcal{F}}$. Given a margin $\alpha \in [0, 1]$, this loss can be written as:

$$\mathcal{L}_m = max(M(\tilde{\mathcal{G}}_{\mathcal{F}}, \mathcal{F}) - M(\mathcal{G}_{\mathcal{F}}, \mathcal{F}) + \alpha, 0).$$

Our second setting (`Aff`), keeps the same model described above, but changes the training procedure so that the inter-atomic term of the docking score becomes the prediction target. However, docking distributions are also unbalanced with only a few high-scoring compounds, that we are most interested about. Thus, we normalize our docking scores by using a quantile transform with 50 bins on each pocket, resulting in values between zero and one. We then train our network to approximate those values using an $\mathcal{L}_2$ loss.

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[32]. 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[11], 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[20] 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`.