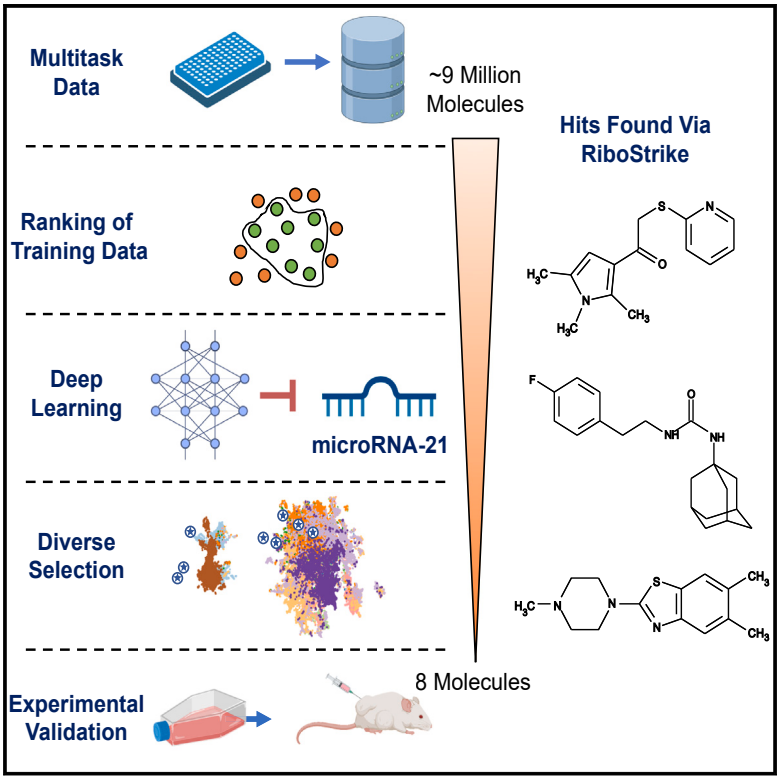
Patterns

Functional microRNA-targeting drug discovery by graph-based deep learning

Graphical abstract



Highlights

- RiboStrike is an AI framework to identify small molecules targeting miRNA-21 activity
- RiboStrike faithfully identifies three candidates with anti-miR-21 activity
- Phenotypic and molecular profiling demonstrates Ribo21-D1's miR-21 selectivity

Authors

Arash Keshavarzi Arshadi, Milad Salem, Heather Karner, Kristle Garcia, Abolfazl Arab, Jiann Shiun Yuan, Hani Goodarzi

Correspondence

hani.goodarzi@ucsf.edu

In brief

This study presents RiboStrike, a deep-learning platform that identifies small molecules targeting microRNAs, specifically miR-21, a known driver of breast cancer. From an extensive screening of nine million molecules and ensuring specificity, eight were identified, with three showing promising anti-miR-21 activity in both reporter assays and RNA sequencing experiments. The potential of these findings is underscored by a significant reduction in lung metastases in a breast cancer mouse model, marking a notable advancement in targeted cancer therapy.

Keshavarzi Arshadi et al., 2024, Patterns *5*, 100909
January 12, 2024 © 2023 The Authors.
https://doi.org/10.1016/j.patter.2023.100909

CellPress

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

Patterns

CellPress
OPEN ACCESS

Article

Functional microRNA-targeting drug discovery by graph-based deep learning

Arash Keshavarzi Arshadi,[1,2,3,4,6] Milad Salem,[5,6] Heather Karner,[1,2,3,4,6] Kristle Garcia,[1,2,3,4] Abolfazl Arab,[1,2,3,4] Jiann Shiun Yuan,[5] and Hani Goodarzi[1,2,3,4,7,*]

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA
[2]Department of Urology, University of California, San Francisco, San Francisco, CA, USA
[3]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA
[4]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
[5]Department of Computer Engineering, University of Central Florida, Orlando, FL, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: hani.goodarzi@ucsf.edu
https://doi.org/10.1016/j.patter.2023.100909

THE BIGGER PICTURE RiboStrike is a deep-learning framework, or a type of machine learning, that navigates the vast chemical space to identify molecules capable of modulating microRNA activity. MicroRNAs are short, single-stranded RNA molecules that modulate different cellular processes, including gene expression, and are implicated in several diseases, such as cancer. The traditional approaches to therapeutically target microRNAs in cancer usually face several experimental limitations. Deep-learning-based strategies, such as RiboStrike, represent a significant change from these more traditional methods and a stride forward to innovations where machine learning assists in unraveling the complexities of disease mechanisms and discovering effective and targeted treatment strategies.

SUMMARY

MicroRNAs are recognized as key drivers in many cancers but targeting them with small molecules remains a challenge. We present RiboStrike, a deep-learning framework that identifies small molecules against specific microRNAs. To demonstrate its capabilities, we applied it to microRNA-21 (miR-21), a known driver of breast cancer. To ensure selectivity toward miR-21, we performed counter-screens against miR-122 and DICER. Auxiliary models were used to evaluate toxicity and rank the candidates. Learning from various datasets, we screened a pool of nine million molecules and identified eight, three of which showed anti-miR-21 activity in both reporter assays and RNA sequencing experiments. Target selectivity of these compounds was assessed using microRNA profiling and RNA sequencing analysis. The top candidate was tested in a xenograft mouse model of breast cancer metastasis, demonstrating a significant reduction in lung metastases. These results demonstrate RiboStrike's ability to nominate compounds that target the activity of miRNAs in cancer.

INTRODUCTION

As a class of short non-coding RNAs, microRNAs (miRNAs) are among the essential regulators of cellular homeostasis. They oversee gene expression and regulate protein synthesis across many target regulons. The dysregulation of these post-transcriptional regulatory programs has been shown to contribute to tumor formation and progression.[1] Since miRNAs regulate a variety of gene regulatory programs, they play an important role in the emergence of various oncogenic hallmarks, such as metastasis,[2] angiogenesis,[3] and resistance to apoptosis.[4] There is recent evidence that they may contribute to the suppression of the immune response within the tumor microenvironment as well. It has also been established that miRNAs are involved in drug resistance in multiple cancers.[5] Among miRNAs, miRNA-21 (miR-21) has been one of the most well-studied drivers of oncogenesis.[6] miR-21 dysregulation has been implicated in ovary, gallbladder, colorectal, pancreatic, and many other cancers.

Much effort has been dedicated to targeting oncogenic miRNAs such as miR-21 to combat various tumors.[7] This is largely achieved through the synthesis and delivery of antisense oligonucleotides (ASOs). For example, researchers developed ASOs that are capable of targeting miR-155, which is known to

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

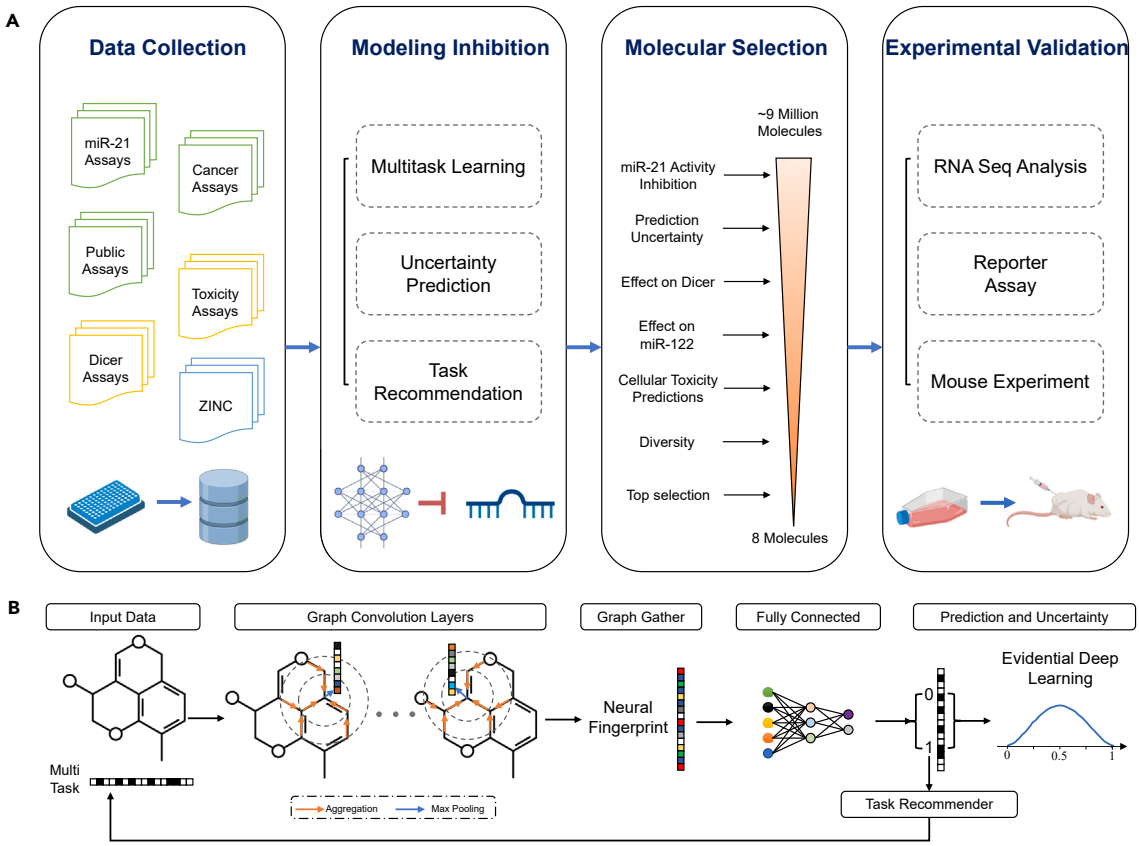


Figure 1. Overview of the RiboStrike pipeline

(A) Stages of discovery from input molecular data and deep-learning techniques to candidate selection and experimental validation. Molecular data are collected from multiple sources for training virtual screening models in multitask learning mode where different datasets are grouped together and share learned representations. A task recommender algorithm helps choose the grouping of the tasks for multitask learning to maximize performance. Nine million candidate molecules are filtered based on the predictions of the models on bioactivity, interactions, and toxicity and the uncertainty in those predictions. After clustering for a diverse selection, eight of the top candidates are experimentally validated.

(B) Computational pipeline and the flow of data within the GCNN network. Molecules are represented as graphs with calculated node and bond features. The convolution layers learn distinctive representations, which are pooled into fixed-length vectors and used for classification by the fully connected network. Once a multitask learning model is trained, an in-house algorithm is used to recommend task grouping for another round of multitask learning.

be involved in multiple types of cancer, including breast and lung tumors.[8] However, delivering ASOs to cancer cells is often challenging due to their poor permeability, and short ASOs may also bind to other RNAs with similar sequences, leading to unintended off-target effects. To overcome these limitations, some researchers have focused on developing small molecules that bind and inhibit miRNAs.[9,10] Unlike ASOs, small molecules are typically easier to formulate and have better bioavailability, making them more suitable for drug development. In recent years, researchers have made progress in developing small molecules that target various types of RNAs, both coding and non-coding.[9] Dovitinib, for example, is a small molecule in early stages of discovery for its ability to treat triple-negative breast cancer and Alport syndrome by inducing the degradation of miR-21 via RNase L recruitment. As another example, risdiplam is an US Food and Drug Administration (FDA)-approved drug for its ability to treat spinal muscular atrophy (SMA) by targeting SMN2 pre-mRNA exon 7-intron junction.[9]

It has been challenging, however, to structurally inhibit the activity of RNAs with small molecules, especially small RNAs such as miRNAs. Unlike proteins, they typically have a dynamic structure[11] and often lack the canonical pockets found in druggable proteins. Therefore, conventional docking approaches have not been successful for miRNAs.[11] Furthermore, structurally binding a small molecule to miRNAs does not necessarily impair their functionality.[12] Another approach to targeting miRNAs is to inhibit their upstream regulators, but outside of the miRNA biogenesis pathway, which is shared among all miRNAs, selective regulators of miRNA activity are largely unknown and poorly characterized. Despite significant research efforts, miRNAs are still considered to be largely undruggable,[13] and a platform that enables selective inhibition of miRNAs remains an important problem in the field.

In this study, we use artificial intelligence to develop a small molecule drug discovery platform called RiboStrike, which aims to inhibit miRNA activity rather than disrupt their structure. Our approach is built on the capabilities of advanced deep-learning architectures to learn representations from molecular data and discover hidden patterns in an abstract and non-linear manner. As shown in Figure 1, we used graph convolutional

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

Table 1. Summary of the datasets

Dataset	Type	Tasks	# Molecules	Active (%)
miR-21[17]	VS	1	315.16K	4.29
Cancer	VS	48	535.45K	2.54
PCBA[18]	VS	128	439K	1.39
Combined	VS	139	540.42K	2.39
Recommended	VS	7	448.87K	4.53
Toxicity[20]	AUX	58	8.36K	18.73
Toxicity recommended	AUX	5	8.36K	19.64
DICER[21]	Off	1	46.72K	6.05
ZINC[16]	MS	–	9.20M	–
Asinex[19]	MS	–	3,652	–
FDA approved[22]	MS	–	157	–
SAR sample[23]	MS	–	37	–

Datasets used for virtual screening (VS), auxiliary modeling (AUX), and molecule selection (MS) as well as the number of tasks (assays), number of molecules, and the percentage of active molecules within each dataset. K, thousands; M, millions.

neural networks (GCNNs)[14] to aid the virtual screening of small molecules against the oncogenic miR-21 *in silico*, solely relying on the simplified molecular-input line-entry system (SMILES) encoding of small molecules as input. We used multitask learning to learn the relationship between chemical groups and molecular activity across multiple data sources, including large publicly available assays from PubChem.[15] Furthermore, we developed several additional modules to improve our model's performance and utility, such as uncertainty prediction, auxiliary modeling (e.g., toxicity prediction), and molecular diversification, to help prioritize molecules for wet-lab follow-up experimentation and validation.

As we demonstrate, RiboStrike selected eight candidate compounds from a pool of more than nine million molecules in the ZINC15 database.[16] We then performed a battery of functional and molecular experiments, in cell culture and in mice, to functionally validate the anti-miRNA activity of our hit compounds. Taken together, our results indicate that RiboStrike can identify candidate inhibitory molecules without the need for the structural information of the target. Additionally, in this work, we have introduced a prediction-based algorithm for recommending input datasets for multitask learning in order to improve the performance of the modeling for the discovery of molecules that inhibit miR-21 activity.

RESULTS

Aggregating data for miR-21, cancer, and off-target interactions

We used three different data modalities to train RiboStrike for optimized hit selection. First, we used datasets relevant to the virtual screening task to train models that predict the effect of a particular small molecule on the activity of miR-21. In this category, we combined nearly 14,000 inhibitors of miR-21 activity (measured by a reporter assay in a 315,000-compound library),[17] assays designed to identify cancer-fighting candidates (cell-based and biochemical assays), and the PCBA dataset (a collection of overlapping PubChem assays[18]). Combining all these datasets resulted in a total of 139 classification tasks; however, in order to create a more focused training dataset that is optimal for multitask learning, we used a supervised task recommender algorithm to narrow down the output of our model to seven tasks. Additionally, we also took advantage of a dataset that measures DICER inhibition to perform an *in silico* counter-screening. DICER is one of the main processing enzymes for miRNAs and its inclusion allowed us to select against inhibitors of DICER activity and identify miR-21 selective inhibitors as opposed to systemic inhibitors of miRNA biogenesis. This model, alongside a model trained on general cellular toxicity, guides the pipeline's prediction to maximize on-target efficacy and specificity and filter overtly toxic compounds. Finally, we used large-scale libraries of drug-like *in silico* compounds as inference datasets to find candidates for further screening. These datasets include ZINC[16] for its diverse and large collection of molecules and Asinex[19] for its selection of molecules (including potential RNA binders). A summary of all datasets used for training, inference, and filtering in this research is listed in Table 1. As mentioned above, our model receives SMILES strings as input, in canonical form with isomeric information included. As for the outputs, each instance carries multiple binary labels denoting the existence of bioactivity or property classes for each of the output tasks.

Computational pipeline: Training GCNNs on small molecules

In learning from molecular data, RiboStrike utilizes GCNNs as its primary modeling approach. Graph-based models are well suited for handling small molecules since nodes represent atoms within molecules and edges represent bonds between them.[14,24] In this graph representation, each node contains features that describe the atom and its properties. Through the use of these features during training, abstract representations of the nodes can be formed by applying graph convolution, which can then be pooled into a representation for the given molecule through the graph gathering layer. We and others have used GCNNs to effectively learn the relationship between input molecules and their physicochemical properties,[25] their impact on a given target's function,[26] and virtual screening.[27] Other models, such as pretrained transformers (e.g., ChemBERTa[28]), which rely on masked token prediction from SMILES sequence or tensor field networks[29] and extract features from coordinate clouds, were also considered for this work. However, GCNNs were ultimately chosen due to their simplicity, smaller size, and training efficiency (empirical comparison of performance to ChemBERTa can be found in Table S2). In this study, we have also taken advantage of evidential deep learning[30] to provide an estimate of the uncertainty of a prediction given an input molecule, which shows the confidence of the model and guides the molecule selection process to have less uncertainty. As shown in Figure 1, we trained GCNN models to predict miR-21 activity suppression, DICER inhibition, and toxicity profiles of each input molecule, given their canonical SMILES. We have illustrated our pipeline for the flow of data in this study in Figure 1B.

Task recommendation: Tailoring the training dataset for miR-21

The high number of tasks that results from combining all assays often degrades the model's performance. A particular problem often arises during multitask learning and stochastic gradient

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

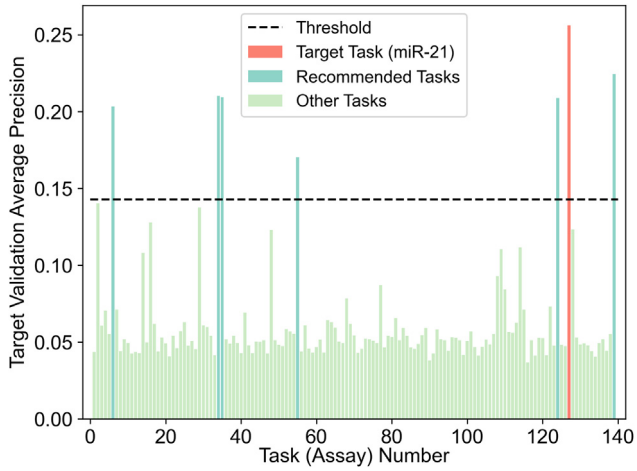


Figure 2. The AP score of different sub-models for the task recommender algorithm

The tasks above the threshold line make predictions matching the miR-21 ground truth to a higher degree than the rest of the tasks. These tasks are the recommended tasks and are selected for training a new multitask model with narrower scope.

descent (SGD) when training on multiple tasks may reduce the performance on one or more of the tasks (e.g., some tasks differ from others regarding the gradient direction), which results in a less efficient training process.[31] To address this challenge in a principled way, we implemented a prediction-based recommendation algorithm to select a subset of tasks from the entire dataset and to create a smaller dataset that is optimized for higher performance for the specific tasks that are of our interest. Using this algorithm, sub-models with similar predictions to the target task (for example, miR-21) are identified, and the top-ranking tasks relating to these sub-models are selected as recommendations. Figure 2 provides an overview of the scores for all sub-models on the target miR-21 task as well as the threshold line we selected. Using our recommender technique, we identified seven tasks that scored higher than the threshold of the mean plus two standard deviations (the tasks are identified in Table S2).

As expected, the counter-screen assay for miR-21 activity,[32] which was an assay to detect false positives in the main screen by measuring the activator of the firefly luciferase, is positioned as one of the recommended tasks. This is intuitive due to this counter-screen's direct association with the main task and the importance of molecular patterns and activity labels contained in this dataset. Interestingly, the remaining recommended tasks were not directly related to miR-21 and were included solely based on the observation that the output of their trained sub-model is similar to that of the miR-21 sub-model. In other words, the models trained on these recommended tasks perform better in predicting the effect on miR-21 activity than the rest of the sub-models, making their respective training data suitable for use in the recommended model.

Virtual screening results: Task recommendation offers the best performance

Following identification of the optimized tasks and implementation of all training scenarios, we compared a variety of modeling techniques to ascertain which dataset and training regimen resulted in the best-performing model. The results on the test set, 10% of the original miR-21 dataset that was held out and isolated throughout model training, were calculated and summarized in Table 2. The confusion matrices for three different training scenarios are shown in Figure 3.

Table 2. Performance of different models on the test set of the miR-21 dataset

Model	# Task	ROC-AUC	Precision	AP
Single task	1	0.8170	12.52	23.03
Multitask: cancer	48	0.8239	11.10	20.62
Multitask: all	139	0.8325	10.14	20.43
Random selected tasks	7	0.8027	18.11	21.04
Recommended tasks	7	0.8305	20.93	24.72*

Models have different types of training data, yet all contain the same miR-21 task for fair comparison. AP score is the main metric for performance comparison. Asterisk indicates best performance.

As was expected, the prediction-based task recommender algorithm resulted in the highest-performing model using the recommended tasks, compared to random tasks, all tasks, or the single miR-21 task, as the model trained on the recommended tasks achieves the highest average precision score. Since our model is conservative, it has a lower tendency to predict molecules as active, and therefore shows a lower performance in recall (in comparison to Figures 3A and 3B). It is evident, however, from the precision score in Table 2 and the confusion matrix in Figure 3C that molecules that are predicted to be active are more likely to be true positives, which is a desirable behavior for virtual screening models. This is because follow-up experimental validations are often costly and we therefore have a higher tolerance for false negatives than false positives in virtual screens.

Evaluation of the RiboStrike model on held-out datasets

To independently verify the performance of our best-performing model, we took advantage of data from an independent study, namely the structural activity relationship sample dataset,[23] which included 37 previously unseen molecules derived from two inhibitors of miR-21. Our virtual screening model successfully classified 33 out of the 37 molecules as "active," closely matching the results of the related study,[23] demonstrating the potential for this model in the context of structural activity relationship (SAR) scenarios. For the four misclassified molecules, our model was uncertain (with uncertainty of 100% for all four misclassified cases). This is likely because, for the most part, these molecules were in the last iteration of SAR and were therefore significantly altered. Consequently, these molecules are outside the familiarity zone of our model's training set, resulting in an uncertain prediction for the model.

Auxiliary models for selecting molecule candidate: DICER inhibition modeling results

As mentioned earlier, miRNAs are transcribed and matured through a predefined pathway. DICER is the main processing enzyme for miRNA biogenesis and inhibiting its activity will reduce the activity of all miRNAs and not just miR-21. To ensure that our model is resistant to this possibility, we added a counter-screening

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

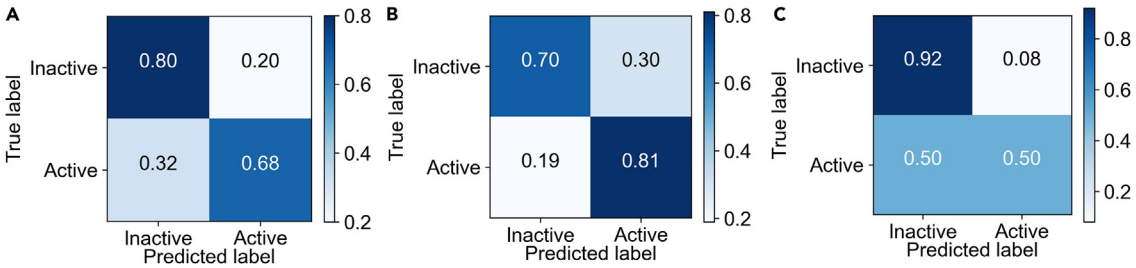


Figure 3. Confusion matrix for models trained using different learning methods

(A) Single task, (B) multitask for all tasks, and (C) multitask for recommended tasks. Balance between true positives (predicted correctly as positive) and false negatives (predicted incorrectly as positive) is needed for virtual screening. Due to imbalanced data, false negatives hurt candidate pool quality drastically.

against DICER to ensure miR-21 specificity of the model.[33] For this, we trained a specialized model on data from an assay regarding DICER-mediated maturation of pre-miRNA[34] to predict inhibitory activities against the DICER as a way to identify and avoid unwanted inhibitory effects. The performance of this model is shown in Table 3.

Auxiliary models for selecting molecule candidate: Toxicity modeling results

It is imperative that candidates for drug development are non-toxic and have as few off-target interactions as possible. As a way to filter any unwanted inhibitory effects on selected targets that could lead to cell toxicity, we trained two toxicity models. In order to determine molecular toxicity against cell viability, we used data from HepG2, a liver cell line that is used as a standard model. Moreover, we trained additional auxiliary GCNN models on the Tox21 dataset,[20] which includes 58 different toxicity tasks, with the purpose of filtering compounds that may be broadly toxic to cells. Overall, the molecules with few toxicity predictions (out of 58) and the lowest uncertainty on HepG2 toxicity prediction pass satisfy this filter. The performance of these two models is described in Table 3. As shown in Table 3, the model trained on the recommended tasks is capable of predicting HepG2 toxicity with a higher average precision (AP). This model is used to predict HepG2 toxicity and uncertainty, whereas the multitask model is used to perform the remainder of the toxicity predictions.

Selecting diverse candidates by clustering the learned small molecule embeddings

Once the virtual screening model and the auxiliary models have been trained, they can be used to screen unseen molecules for their potential as drug candidates. Given its large and diverse collection of drug-like and synthesizable compounds, we used the ZINC15 library for our virtual screening; however, we also included compounds from the Asinex library, which represents an out-of-distribution collection and allows us to assess the generalizability of our model. We used our trained model to screen these datasets for suitable molecules that are most likely to specifically inhibit miR-21 activity with fewer potential side effects and least likely to cause overt toxicity. We first used our multitask virtual screening model trained on the seven recommended tasks to predict miR-21 activity inhibition across nine million molecules *in silico*.

To obtain a more complete understanding of how diverse the inference molecules are, we performed unsupervised analysis and clustering of the molecular embeddings learned by the model. This analysis also enabled us to select molecules from various regions of the molecular space. To accomplish this, the embedding features of the trained model were extracted, projected into a two-dimensional (2D) uniform manifold approximation and projection (UMAP) space for visualization, and clustered using the k-means algorithm. By separating molecules into clusters, different regions of the molecular space can be accessed to select more diverse molecules for follow-up and testing (Figure 4). In Figure 4A, this molecular space is depicted using the positively predicted candidates from the ZINC database, which occupies a wide range of the embedding space. Figure 4B depicts the same embedding space but for the molecules used for model training, showing the overlap between the positive and negative compounds in the training set. In Figure 4C, we have shown clusters within the feature space as well as the top eight compounds that were selected from these clusters using the RiboStrike algorithm.

Assessment of the learned small-molecule embeddings and selection of hit compounds for further validation

It is evident from Figure 4A that the molecules from the ZINC database that have been predicted to be positive cover most of the projected space, demonstrating that the model is not learning a singular molecular feature. It is also apparent that FDA-approved drugs[22] are concentrated in certain clusters, which likely makes the selection of hit molecules from their vicinity favorable. Moreover, the positive and negative training data distributions in Figure 4B have a substantial overlap, which highlights the challenge in virtual screening for miR-21. By clustering the hit molecules into 10 groups across the embedding space, we ensured that our selected hit compounds were sampled from varying clusters and therefore retained molecular diversity.

Table 3. Performance of toxicity and DICER inhibition models on their corresponding test sets

Model	# Task	ROC-AUC	Precision	AP
DICER single task	1	0.6631	12.22	12.53*
Tox multitask	58	0.7408	41.13	38.41
Tox recommended	5	0.7483	38.27	38.73*

The models belonging to the toxicity category are comparable. Asterisks indicate best performance.

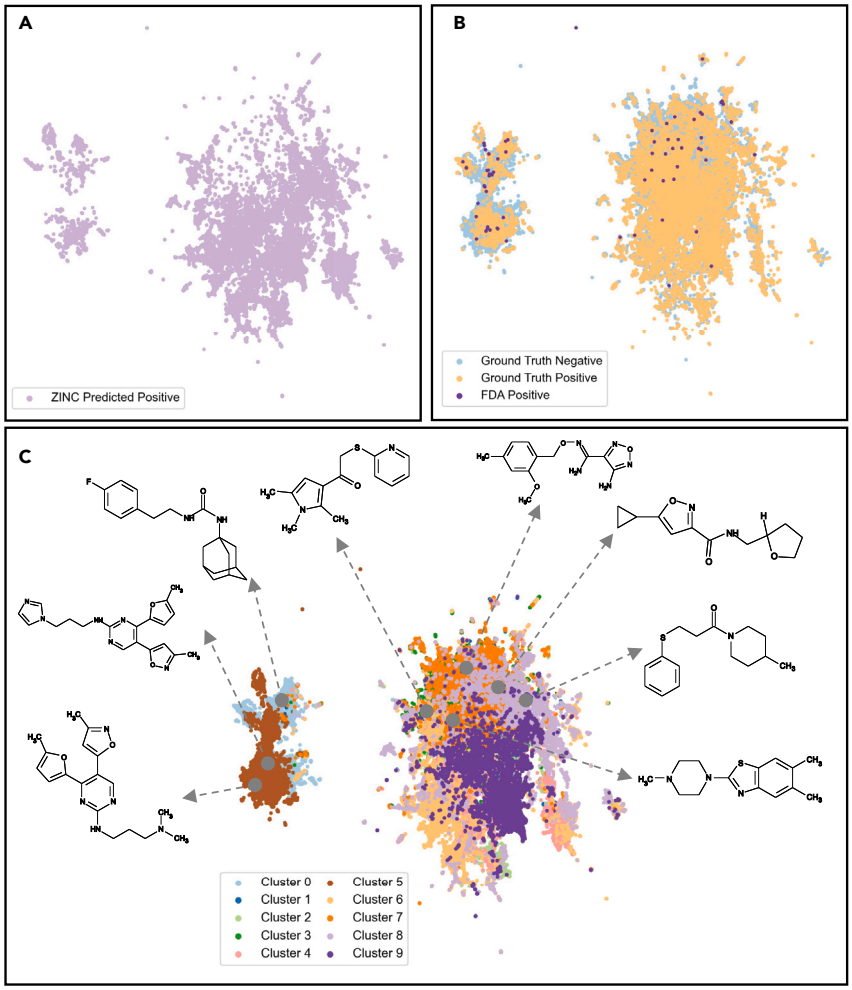


Figure 4. UMAP of the inner features of the model for the inference and training data for

(A) ZINC data predicted to be active as the candidate space to select from, (B) the training data and the positive molecules from the FDA-approved set, showing that the candidate space spans the same area as the training set, and (C) the 10 clusters applied to this space for the inference sets (ZINC and Asinex) and the final selected molecules from these clusters to ensure diversity in selection.

Using clustering as the basis for the selection of molecules, those hit compounds that are predicted to affect the activity of miR-21 with high confidence were selected and virtually screened against DICER inhibition and cellular toxicity. To conclude, we selected a total of 10 molecules across the two inference datasets (ZINC and Asinex) specifically from embedding clusters where FDA-approved molecules are well represented. We also required them to have low uncertainty in predicted inhibition of miR-21's activity, with few or no toxic activity predictions and without predicted activity against DICER. Overall, six molecules were selected from the ZINC dataset and four from Asinex. Of these, we successfully acquired eight molecules, six from ZINC and two from Asinex. We ensured that none of these molecules had previously been studied in this context and by and large represented new chemical entities. The selected molecules are identified by their IDs and SMILES in Table S3.

RiboStrike training on miR-122 data for generalization and counter-screening purposes

While we had focused the bulk of our study to training RiboStrike against the oncogenic miR-21, the platform itself is general in concept and can be applied to other miRNAs or even other RNAs, for which activity can be measured in scalable formats. To demonstrate this possibility, we trained RiboStrike using an NIH dataset that measured miR-122 activity using a reporter system, similar to the dataset we had used for miR-21. As shown in Figure S4, RiboStrike achieved a performance comparable to the one we achieved for miR-21, with regard to recall and true-negative rate. In addition to establishing the possibility of extending RiboStrike to other miRNA targets, this model also allowed us to assess the target selectivity of our predictions between the miR-21 and miR-122 models This counter-screen enabled us to confirm that the selected miR-21-inhibitor molecules were specific to miR-21 and did not affect other miRNAs *in silico* (Table S4).

Gene expression profiling to measure miR-21 activity in response to treatment with the selected hit compounds

Since miR-21 is a post-transcriptional regulator of RNA stability, reducing its activity results in an increase in the RNA levels of its target regulon. To experimentally verify the anti-miR-21 activity of our selected compounds, we used an RNA sequencing (RNA-seq) strategy amenable to scalable gene expression profiling, namely QuantSeq-Pool. For experimental testing, we used MDA-MB-231 cells, an established model of triple-negative breast cancer metastasis that is known to be driven by miR-21.[34] Inhibition of miR-21 in these cells should significantly reduce their metastatic potential. We first used CellTiter-Glo to calculate the IC20 (inhibitory concentration of 20%) for each of our compounds, to ensure a regimen in which the key cellular processes are not affected by each treatment. We then performed QuantSeq-Pool on MDA-MB-231 cells treated at IC20 for 72 h in biological replicates. We also included DMSO-treated control samples. For positive control, we used an established anti-miR-21 ASO and included a non-targeting ASO as control. Upon measuring the gene expression changes induced by each compound, we asked whether they caused a systematic effect on the expression of miR-21 target RNAs. We used the set of RNAs that are annotated as miR-21 targets (based on Targetscan[35]) to perform gene set enrichment analyses (Figure S3). As expected, in the miR-21 ASO samples, we observed a significant enrichment of miR-21 targets among genes that were upregulated upon treatment.

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

From the eight selected compounds, two show miR-21 target upregulation similar to that of the ASO, and a total of five show some activity (with a confidence score of over 85%). Based on these results, our model has a hit rate of 62.5%, higher than the *in silico* predicted precision score of 20.93% for our model.

Reporter assays for targeted measurement of miR-21 activity in dose-response assays

For the five compounds showing anti-miR-21 activity based on RNA-seq data, we also used a reporter assay for an independent confirmation at multiple doses. For this, we designed a GFP reporter harboring two miR-21 recognition sites in its 3′ untranslated region (UTR) and transfected it into MDA-MD-231 breast cancer cells. This construct also drives the expression of an mCherry gene, which serves as an endogenous control for the construct.[36] Through the use of miR-21 ASOs, non-targeting ASOs, and flow cytometry, we confirmed that inhibiting miR-21 resulted in a significant increase in GFP expression in these reporter cell lines. We then performed this experiment for each compound separately. As shown in Figure 5A, three (Ribo21D-1, Ribo21D-2, and Ribo21D-3) of the five compounds also showed significant dose-dependent activity against miR-21. Taken together, our platform showed an experimentally confirmed hit rate of 37.5% for entirely new chemical entities that were predicted *in silico*.

Total RNA and small RNA-seq for deeper analysis of the top three anti-miR21 compounds

To perform a high-resolution analysis of gene expression changes, we performed total RNA-seq for MDA-MB-231 cells treated with the three validated compounds (Ribo21D-1, Ribo21D-2, and Ribo21D-3) as well as DMSO control cells, in biological triplicates. We also performed RNA-seq for cells transfected with anti-miR21 ASOs and non-targeting controls (also in triplicates). First to confirm our results from QuantSeq-Pool and the ability of these compounds to decrease miR-21 activity, we defined a putative miR-21 target regulon by combining predictions from TargetScan[35] and miR-DB.[38] We then assessed the changes in the expression of this putative target regulon across each treatment. We performed this analysis in two ways: first, we performed the gene set enrichment analysis we have previously described,[39] where the genes are sorted based on their log fold changes in each treatment and divided into equally populated bins. The enrichment or depletion of the target regulon is then assessed and visualized for each expression bin. The resulting heatmap showed an expected depletion in the leftmost bins (i.e., genes with lower expression upon miR-21 inhibition) and significant enrichment in the rightmost bins (i.e., genes with increased expression). The overall statistical association was measured using mutual information and the associated Z score (Figure 5B, left). In all three cases, we observed a significant depletion and enrichment pattern, mirroring the pattern observed for the ASO positive control. As a complementary analysis, we also performed t test between the log fold change (logFC) values for the putative miR-21 targets and the rest of the transcriptome. Consistent with our gene set enrichment analysis, all three treatments significantly increased the expression of this putative regulon above background (Figure 5B, right).

To ensure that our results were not biased by our selection of putative miR-21 targets, we also assembled an empirical miR-21 target regulon. For this, we intersected mRNAs bound by RNA-induced silencing complex (RISC) complex loaded with miR-21 based on the microCLIP dataset[37] with those upregulated in our cells upon transfection with the anti-miR21 ASO. Similar to the putative regulon, this empirical miR-21 target set showed a similarly significant increase in their expression in each of our treatments (Figure 5C). Together, our results confirm the role of our hit compounds in increasing the expression of miR-21 regulon, which is consistent with the lower activity of this miRNA.

These compounds may be affecting miR-21 expression or activity. In the former, we expect reduced miR-21 levels due to a reduction in its biogenesis, whereas, in the latter mechanism, miR-21 levels do not change. The anti-miR21 ASOs are an example of this mechanism as they bind and sequester miR-21 and reduce its activity but not expression. To determine whether these compounds are working upstream or downstream of miR-21 and whether their effect is specific to miR-21, we performed small RNA-seq to profile miRNAs upon treatment of MDA-MB-231 cells in biological triplicates. As shown in Figure 5D, two out of three compounds, namely Ribo21D-1 and Ribo21D-3, resulted in a significant reduction in miR-21. However, a broader look at all miRNAs revealed that Ribo21D-3 resulted in a significant reduction in the expression of multiple additional miRNAs (Figure S3). Ribo21D-1, however, proved to be quite selective against miR-21, as no other miRNA was observed to be downregulated upon treatment. We did observe higher expression of four miRNAs, namely miR-663, miR-3196, miR-1908, and miR-941; however, the higher expression of these miRNAs is downstream of miR-21 downregulation. This is because these miRNAs are also upregulated upon anti-miR21 transfection, resulting in logFC values of 0.7, 1.1, 1.3, and 1.35 for these miRNAs, respectively. Together, our results confirm the anti-miR-21 activity of our selected compounds and highlights the fact that these compounds likely function through independent mechanisms. Ribo21D-1 is a selective inhibitor of miR-21 biogenesis, Ribo21D-2 acts downstream of miR-21 modifying its activity, and Ribo21D-3 is a non-specific regulator of miRNA biogenesis and lowers the expression of several miRNAs. This diversity in mechanism of action is consistent with our effort in selecting diverse compounds from our virtual screening hits.

Measurement of anti-metastatic activity in xenograft mouse models

As mentioned earlier, miR-21 is a driver of metastasis in breast cancer. Therefore, we expect the inhibition of this miRNA to result in lower metastatic potential in breast cancer. To confirm this, we treated MDA-MB-231 breast cancer cells for 72h with the most promising candidate molecule, Ribo21D-1. We then injected these cells, along with DMSO-treated controls, into immunocompromised NOD-SCID gamma (NSG) mice via their tail-veins. We used *in vivo* imaging to monitor the colonization and growth of cancer cells in the lungs of mice over a period of roughly 40 days. As shown in Figure 5E, we observed that, consistent with anti-miR-21 activity, pre-treating breast cancer cells with Ribo21D-1 resulted in a significant reduction in their lung colonization capacity. Therefore, this compound not only reduced miR-21 expression and activity, as measured by small RNA-seq, RNA-seq, and reporter assays, but also reduced metastatic lung colonization in xenografted mice (Figure 5E).

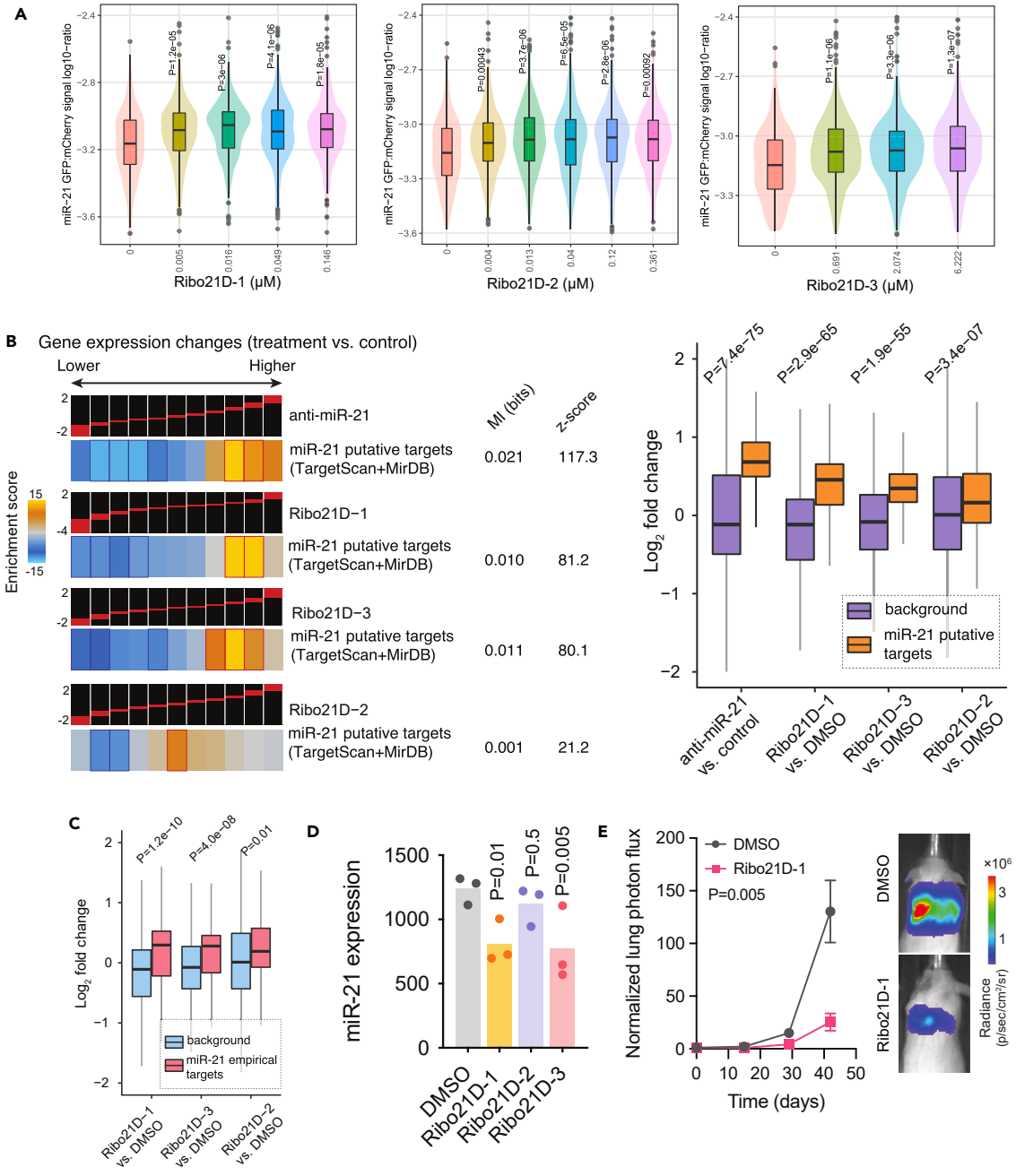


Figure 5. Experimental validation results

(A) Dosage response assay in reporter cell lines for Ribo21D-1, Ribo21D-2, and Ribo21D-3.

(B) Enrichment and depletion patterns of putative miR-21 targets across gene expression changes for anti-miR21 as well as Ribo21D treatments (left). For this analysis, gene expression changes (log2 fold changes) were first sorted and divided into equally populated bins, from downregulation (left) to upregulation (right). For each analysis, the mutual information value (in bits) and its associated *Z* score are provided. The heatmaps show the enrichment and depletion patterns from the onePAGE package. Blue marks depletion and gold marks enrichment. Expression bins with statistically significant depletion or enrichment of miR-21 putative targets are marked by a dark blue or red border. As an alternative visualization, we have also shown the logFC values for the miR-21 targets and the background set of genes (right). The p values were calculated using t test.

(C) Expression of empirically determined miR-21 targets (microCLIP[37] plus anti-miR transfections) in response to the top three compounds.

(D) Expression of miR-21 determined using small RNA-seq across the three treatments. The p values were calculated using DESeq2.

(E) *In vivo* lung colonization assays were used to measure impact of Ribo21D-1 on lung metastasis. Normalized lung photon flux, which measures luciferase activity in labeled cancer cells, as a function of time for each cohort (n = 5) is shown. Two-way ANOVA and Mann-Whitney U tests were used to assess statistical significance. Also shown is a representative mouse and lung image.

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

CellPress
OPEN ACCESS

Together, our findings establish the utility of RiboStrike as an effective platform for discovering compounds against miRNAs.

DISCUSSION

In light of the dynamic structure and small size of miRNAs, the discovery of inhibitory small molecules against them presents a number of challenges. This is further compounded by the fact that miRNA ligands do not necessarily interfere with the function and activity of miRNAs. Considering this, we propose discovering candidate hits that instead focus on targeting miRNA activity. The goal of this study was to implement a virtual screening platform that leverages deep learning to enable the selection of early hit candidates out of a large collection of diverse molecules. Multiple methods were implemented in order to ensure the practicality of the computational methods, as well as rigorous experimental characterization, resulting in the validation of multiple compounds *in vitro* and for our top hit also *in vivo*. The first step involved the use of deep-learning models to train on a large number of small molecule datasets to learn the chemical language underlying miR-21 activity. We also introduced a new task recommendation technique, which identified the optimal configuration for combining datasets to maximize training potential. The third methodology involved calculating uncertainty for all predictions made by the models, which enabled the ranking of molecules in spite of their binary nature. To finalize, the internal features of the model were used to represent molecules during inference, and clustering of these embedding features allowed the selection of a diverse set of molecules for experimental testing. Finally, we conducted additional training of the RiboStrike model using miR122 inhibitor data. The objective of this model was to filter potential miR-21 activity inhibitors and ensure the selectivity of the chosen molecules. It is crucial to note that the primary aim of this project was not to identify molecular binders of miR-21 but rather to uncover molecules capable of inhibiting miR-21 activity. Consequently, these molecules could potentially bind to an unidentified regulator of miR-21 to produce the desired effect. In total, the RiboStrike platform identified multiple hit candidates that were subsequently confirmed, demonstrating the advantage of using graph-based deep learning to identify hidden patterns of molecular hits against the activity of miRNAs without the need for sequence reading or structural information. The ultimate objective of utilizing miR-21 data is to construct a model capable of identifying molecules exhibiting ASO mimicry, rather than focusing on elucidating the precise molecular mechanisms behind miR-21 inhibition. Additionally, we integrated data from miR-122 and DICER, incorporating biological insights into potential inhibitors with fewer cellular side effects. It is likely that these compounds target molecular pathways upstream of miRNAs that regulate their processing, lifespan, and function. Therefore, once identified and validated, these hits can be used as toy compounds to identify their direct targets (e.g., using CRISPRi screen with miRNA reporter expression as readout).[40] The not only is identification of these direct targets required for additional genetic and biochemical validations of their impact on the activity of specific miRNAs but this knowledge provides an avenue for further optimization of our tool compounds using traditional SAR and medicinal chemistry. Therefore, this approach may provide us with a better understanding of the mechanisms by which miRNAs are regulated through the discovery of crucial unknown players upstream.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Hani Goodarzi (Hani.Goodarzi@ucsf.edu).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

The data and code used in this work are available through our GitHub repository at https://github.com/LumosBio/RiboStrike. The version of this code and data at the time of publication is also captured and saved through Zenodo at https://doi.org/10.5281/zenodo.10068059.[41]

Data sources

Three categories of data are used throughout the RiboStrike pipeline: virtual screening datasets, off-target interaction datasets, and inference datasets, which are explored further in the following subsections. Each dataset is preprocessed to contain canonical SMILES as well as appropriate binary labels for the activity of molecules, and the details of this process can be found in the supplemental information. SMILES is a widely used molecular descriptor that encapsulates the structure of a molecule, serving as an input for models. Graph convolutional neural networks (GCNNs) utilize the SMILES input to create a graph representing the molecule, which in turn is used for feature representation learning. For instance, the SMILES descriptor "Cc1cc(C(=O)CSc2ccccn2)c(C)n1C" represents the molecular structure of Ribo21D-1.

Virtual screening training datasets

The datasets used to train the main virtual screening model or to help its performance via multitask learning are as follows.

- miR-21 Data: primarily, an HTS (high throughput screening) dataset from miR-21 inhibition screening as a target is used in this study, a data repository originating from National Center for Advancing Translational Sciences (NCATS) and deposited in PubChem (ID: AID 2289[17]). The aim of this assay had been the discovery of molecules with an inhibitory effect on miR-21 to finally induce cell apoptosis and tumor suppression. They used a cell-based firefly luciferase reporter gene assay optimized for qHTS (quantitative high throughput screening).
- Cancer-related data: to assist the multitask training process, different cancer-related assays are collected from PubChem and the PCBA dataset. These assays include 20 tasks directly from the PubChem database, and 38 cancer-related tasks from the PubChem BioAssay (PCBA) dataset.[18].
- PCBA dataset: PCBA is a collection of datasets aggregated from PubChem consisting of the biological activities of small molecules generated by high-throughput screening. In this work, a subsection of PCBA with 128 bioassays is used with over 400,000 molecules, similar to the previous benchmarking methods.[18] This dataset was selected due to its size, high number of tasks, and high molecular overlap with the miR-21 dataset. Due to these features, this dataset can be combined with the miR-21 dataset to create a large non-sparse training set for multitask learning.

Two further datasets are created from the mentioned dataset; the "combined" dataset from aggregating all data points, and the "recommended" dataset from algorithmically selecting tasks from the combined dataset using the task recommendation approach (discussed in the "prediction-based task recommendation" section). The preprocessing code for the miR21 data can be found in the GitHub repository under preprocessing_mir21 script. Once preprocessed, these data are augmented with the auxiliary datasets in the preprocessing_merge_pcba script in the GitHub repository.

Off-target interactions datasets: Toxicity and DICER inhibition

To increase the probability of any predicted bioactive molecule to be a drug candidate, the toxicity properties and the effect of these molecules on the DICER's function are predicted. Therefore, two datasets related to these subjects are used for training in this work:

- Toxicity in the twenty-first century (Tox-21): the most popular source available for the analysis of the toxicity of molecules.[20] This dataset consists of 58 distinct tasks, each tested on an important protein target of humans. Candidate drugs that interact with those targets are likely to cause side effects and cellular toxicity. Therefore, we used 58 tasks from tox21 to filter out molecules that would show side effects in future steps.
- DICER dataset: the DICER protein plays an important role in the maturation of several RNAs, not just miR-21. Therefore, if a candidate drug decreases the activity of the miR-21 by inhibiting the DICER protein, it would show significant side effects downstream. The dataset for this task is taken from PubChem data source of AID 1347074.[21] This dataset was created using click chemistry and was selected due to the fact that this assay identified inhibitors of DICER based on pre-miR-21.

The preprocessing code for the auxiliary datasets can be found in the preprocessing_merge_pcba script of the GitHub repository.

Inference data: Sources for drug candidate selections

There are multiple datasets used in this work for the drug candidate selection:

- ZINC: ZINC database is a diverse molecular library typically used for virtual screening.[16] It contains millions of molecules that are indexed for search and properties. We used the drug-like subset (molecular weight: 250–500 g/mol, logP: −1 to 5) of the ZINC15 database, consisting of nine million molecules that had three-dimensional (3D) representations, standard reactivity, reference pH, a charge of −2 to +2, and were in stock for purchase.
- Asinex: Asinex is a molecular library vendor that sells varieties of different classes of molecules, including macrocycles, alpha helix mimetics, and peptidomimetics. This vendor also offers a small-molecule library specifically designed to target RNA,[19] which is a suitable candidate for this project. Moreover, the candidates in this set are different from the training set, allowing for testing the model's generalizability.
- FDA approved: this dataset was taken from a study[22] that found 86 molecules that were suppressors of miR-21 activity from a pool of 696 compounds from the Bioactive Compound Library and 262 compounds from FDA-approved Drug Library. In this work, these molecules are deleted from the mentioned inference datasets, to avoid selection of previously discovered molecules. Moreover, these molecules will be used to assist the selection of the final molecules by identifying the favorable clusters as described in the "molecule selection" section.
- SAR sample: this dataset originates from a study[23] that takes two molecules that are known inhibitors of miR-21 and uses SAR to optimize these molecules and assist in overcoming chemoresistance in renal cell carcinoma. Overall, 37 molecules are tested and shown to be active, which can create a small validation set for this work.

After training, the inference data are loaded and preprocessed with the same steps as the training data. Then the trained model is used to make predictions on the inference sets as well as to attach uncertainty values to these predictions. The inference script within the GitHub repository contains the code for these necessary steps.

Graph convolutional neural network training

In recent years, GCNNs have proved to be helpful in learning representations from small molecules and modeling tasks such as virtual screening,[26] molecular property prediction,[25] and drug-target interaction.[42,43] This success is owed to two facts: first, small molecules are inherently similar to graphs, with atoms represented as nodes and bonds represented as edges, making GCNNs suitable tools for handling this data type. Second, the feature extraction in the GCNN model, which is inspired by traditional circular fingerprint extraction from molecules, results in useful and often superior inner features due to the automatic representation learning aspect of deep learning.[14,24]

In this work, the GCNN implementation from the DeepChem library is used.[44] In this model, the molecules are converted to graphs and atoms, then featurized to include features such as atom type, number of directly bonded neighbors, implicit valence, formal charge, and hybridization type. Isomeric information is also added to the features in the form of a vector with length of three (whether chirality is possible, right-hand, and left-hand). The model is trained in a multitask learning manner, with the intent to share the representation learning mechanism between multiple datasets and increase the efficiency of the training and the performance of the final model. This multitask learning approach is paired with an in-house algorithm that recommends a unique grouping of the tasks in multitask learning to optimize the final performance of the model. The logits of the model are viewed through the lens of evidential deep learning to attach uncertainty values to each prediction, The code used for model training can be found in the training script in the RiboStrike GitHub repository.

The molecular data are split using Murcko scaffolds to create training, validation, and test sets. The hyper-parameters of this model as well as the length of the training are found through hyper-parameter optimization in a grid-search manner via monitoring the performance of the model on the validation set. The code for this optimization can be found in the hyper_opt script within the GitHub repository (more on hyper-parameter optimization can be found in the supplemental information).

Evaluation metric

The metric used for evaluation in this work is the AP score. This metric computes the area under the precision-recall curve and was chosen due its fairness toward imbalanced datasets, where the positive label discovery is of importance. This is the case with most virtual screening tasks, where the number of active molecules is often much lower than the number of inactive molecules, resulting in highly imbalanced datasets. Moreover, discovery of these active candidates is of utmost importance in an early drug-discovery pipeline since these candidates will be passed on to the next steps of the drug discovery process. Therefore, AP score is favored in this work for comparison of models, different architectures, or different epochs during training. The results are also reported for accuracy, recall, precision, and the area under the receiver operator curve (ROC-AUC).

Optimizing the tasks for multitask learning

Multitask learning has proved to be beneficial in many instances via providing multiple tasks for the model to simultaneously learn from, with the hope that the learned representations for these tasks benefit from being shared within the same model. However, this is not the case in all scenarios and, in some cases, negative transfer occurs, where multitask learning hurts the performance of a given task when compared to single-task learning.[31] In this work, to address the problem of negative transfer, the method of prediction-based task recommendation is proposed, which narrows the number of tasks selected for multitask learning via recommending a few tasks in an algorithmic manner.

Prediction-based task recommendation

To begin the process of task recommendation and selection of fewer training tasks, a multitask learning model is trained on all available tasks. After training, one target task is selected (e.g., miR-21 dataset) and inference is performed on the validation set of this target task. Since the model is trained on multiple tasks, it will have multiple predictions for each input molecule, each assigned to one input task. Given N total training tasks and M molecules in the target task's validation set, the predictions of the model will then have a shape of $M \times N$, with each row representing the output of the sub-model assigned to one input task, denoted by $Output_i$. After this output is calculated, each $Task_i$ is scored using the scoring metric in Equation 1.

$$Score_i = AveragePrecisionScore(Label_{target}, Output_i) \quad \text{(Equation 1)}$$

Here, $Label_{target}$ denotes the ground truth labels for the target task. As can be seen from this equation, the labels are kept constant on the target set while the sub-model changes, which is the main difference between this method and

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

simple inference. Using this scoring mechanism, the predictions of different sub-models are compared to the ground truth labels of the target task, with the sub-models that have more similar predictions to the target labels having a higher score. Through the identification of sub-models with similar predictions, this approach identifies their corresponding training tasks and selects the highest-scoring tasks for training. The recommended tasks are selected via applying a threshold of mean plus two standard deviations to the scores. This threshold is arbitrary and can be replaced with a simple selection of top K scores. The recommended tasks are then passed on as training data to the hyper-parameter optimization and training step for the final model. This recommendation process is repeated for the toxicity model as well, with the target task of HepG2. The code for this recommendation algorithm can be found in the training script of the RiboStrike GitHub repository.

Molecule selection: Inference and uncertainty prediction

After different categories of models are trained, the final models are used to predict the properties of the molecules from the inference datasets. All training data, as well as the FDA-approved data,[22] are removed from the inference datasets to avoid the selection of redundant or previously discovered molecules. During the inference process, a binary label is predicted for each given molecule, reflecting its predicted effect on miR-21's function. Having only binary labels to distinguish between different molecules is problematic, since molecules with the same predictions (e.g., active) become indistinguishable and no ranking can be assigned to the molecules for further drug candidate selection.

In order to overcome the selection challenges, an uncertainty prediction method is applied on the last layer of the model. To do so, evidential deep learning is used,[30] which applies a Dirichlet distribution on the class probabilities and computes uncertainty for each prediction. This uncertainty score ranges from zero to one, with lower scores demonstrating more certain predictions. With this uncertainty score, the predictions become distinguishable and the molecules that are predicted to be active with low uncertainty become desirable. The code for this molecule selection process is located within the inference script of the RiboStrike GitHub repository.

Molecule selection: Neural fingerprints clustering

After the uncertainty is predicted, one challenge is faced, which is the lack of diversity in the top selected molecules with low uncertainty. The reason for this phenomenon is that similar molecules result in similar predictions and uncertainty, and the most certain predicted molecules are similar to each other and they populate the top of the certainty ranking list. This creates a problem in the further stages and specifically *in vitro* screening, where diversity among the candidates is needed to increase the chance of activity against different targets.

To enforce diversity within the selected molecules and create variety within the final selection, the molecules are clustered and a few molecules are selected from some of the clusters. To do so, the neural fingerprints of the molecules are extracted from the Graph Gather layer of the trained model. This fingerprint is the inner features of the model for an input molecule and is a numerical vector that can meaningfully represent this molecule. Neural fingerprint clustering allows the molecules belonging to different clusters to both be structurally different and exist in different locations within the feature space of the trained network. After the features are extracted, KMeans (K = 10) clustering is applied to the features, then visualized using 2D UMAP, resulting in 10 clusters formed from the inference molecules. The code for ensuring diversity within the final selection is located under the clustering script within the GitHub repository.

Molecule selection: Five criteria for final selection

After the molecules are clustered and uncertainty and bioactivity are predicted for all of them, five different criteria were checked for the final molecules to be selected:

(1) Potency as miR-21 activity inhibitor: the selected molecules should be predicted to inhibit the function of miR-21.
(2) Certainty: the molecules that had the least uncertainty in each cluster were considered.
(3) Diversity: molecules should belong to different clusters in regard to the clustering of the neural fingerprint. Clusters that include more of the FDA-approved molecules are more likely to be selected from.
(4) Pass majority of toxicity tests: the selected molecules are more likely to be predicted as non-toxic in most of the toxicity tests with low uncertainty for specifically the HepG2 test.
(5) Low chance of interaction with miR-122: the selected molecules are predicted to be specific to miR-21 and do not interact with other miRNA, specifically miR-122.
(6) Low chance of inhibiting the DICER: the selected molecules are more likely to be predicted to not affect the function of the DICER with low uncertainty.

Following these criteria, the inference molecules are first narrowed down to those predicted to be active with high certainty and then filtered via selecting top molecules from each 10 clusters. Afterward, the final molecules are selected from this list with consideration of toxicity and DICER activity and their uncertainties. In the end, eight molecules are selected from the inference datasets (ZINC and Asinex) and progress to the *in vitro* screening stage.

Cell culture

The MDA-MB-231 (MDA-parental, ATCC HTB-26) human breast cancer cell line; its highly metastatic derivative, MDA-LM2[45]; its triple reporter version, MDA-MB-231tr; and HEK293T cells (ATCC CRL-3216) were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (FBS), penicillin, streptomycin, and amphotericin B. Cells were all incubated at 37°C at 5% CO_2 in a humidified incubator.

CellTiter-Glo cell viability assay

MDA-MB-231 cells were seeded at 1,000/well in white opaque 96-well plates (Corning 3917) and treated with serial dilutions of drug candidates ranging from 3.2 pM to 1 mM for 72 h in triplicate. Cell viability was measured using CellTiter-Glo 2.0 Assay (Promega G9243) to find IC20 concentrations.

High-throughput sequencing data generation

MDA-MB-231 cells were seeded at 3,000/well in clear 96-well plates and treated with drug candidates at IC20 for 72 h in duplicate. Controls were transfected with oligo inhibitors targeting either mir-21 or non-targeting controls in duplicate. For transfection of control wells, we added 0.5 μL of 100 μM oligo inhibitor, 100 μL of OptiMEM (Thermo Fisher Scientific 31985088), and 2.5 μL of Lipofectamine 2000 (Thermo Fisher Scientific 11668019), which were incubated together for 20 min at room temperature. Cells were then incubated at 37°C for 72 h. RNA was extracted using the Quick-RNA 96 kit (Zymo Research R1052) and concentrations were determined using Nanodrop. Libraries were prepared using the QuantSeq-Pool Sample-Barcoded 3′mRNA-Seq kit (Lexogen 139) with 10 ng of input RNA. Libraries were sequenced using NovaSeq 6000 SP (100 cycles).

High-throughput sequencing data generation for RNA-seq and smRNA-seq

MDA-MB-231 cells were seeded at 1×10^5/well on 24-well plates and treated with either Ribo21D-1, Ribo21D-2, or Ribo21D-3 at 0.1 μM for 72 h in triplicate. Controls were transfected with oligo inhibitors targeting either mir-21 or non-targeting controls in triplicate. For transfection of control wells, we added 0.5 μL of 100 μM oligo inhibitor, 100 μL of OptiMEM (Thermo Fisher Scientific 31985088), and 2.5 μL of Lipofectamine 2000 (Thermo Fisher Scientific 11668019), which were incubated together for 20 min at room temperature. Cells were then incubated at 37°C for 72 h. RNA was extracted using Direct-zol Mini-prep Kit (Zymo Research R2051). RNA RIN values and concentrations were assessed using the Agilent D1000 ScreenTape System and an Agilent Tapestation 4200 according to manufacturer's instructions. RNA-seq libraries were prepared using the SMARTer Stranded Total RNA-seq V3-Pico Input Mammalian: cDNA kit (Takara 634486) with 5 ng of input RNA. The smRNA-seq (small non-coding RNAs) libraries were prepared using the SMARTer smRNA-Seq Kit (Takara 635031) with 30 ng of input RNA. Libraries were sequenced on a HiSeq4000.

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

QuantSeq-Pool data analysis

The QuantSeq-Pool data were demultiplexed and preprocessed using an implementation of the pipeline provided by Lexogen (https://github.com/Lexogen-Tools/quantseqpool_analysis). The outputs of this step are gene-level counts for all samples. The raw counts matrix used for differential expression analysis was prepared using the DESeq2 package.[46] The logFCs from multiple differential expression comparisons used gene set analysis (Figure 5A).

Finally, we aimed to sort molecules based on their systematic effect on expression of miR-21 target mRNAs through gene set enrichment analysis. Thus, we downloaded TargetScan prediction of hsa-mir-21 3p target genes from miRBase dataset (miRBase: MI0000077). Then, we use a modified version of iPAGE[39] (we call it onePAGE from here on) in which you can perform the gene set enrichment analysis powered by mutual information evaluation and statistical tests for a single gene set. The onePAGE analysis here reports enrichment of miR-21 target gene list in three bins of logFC values; from left to right, (0) lowest bin of log2FC (i.e., downregulated), (1) log2FC around zero (i.e., no expression change), (2) highest bin of log2FC (i.e., upregulated). After running this analysis for differential expression log2FC of all drugs and control conditions, we sorted drugs based on resulted *Z* score keeping miR-21 vs. negative control (nc) at the top (Figure 5A). From this step, we selected the top five drugs for further evaluation.

All scripts for preprocessing, differential expression, and onePAGE enrichment analysis are accessible in this GitHub repository (https://github.com/goodarzilab/targeting-miR-21-RNA-seq).

RNA-seq and small RNA-seq data analysis

For RNA-seq, salmon (v0.14.1) was used to measure gene expression (gencode v44 basic annotation). Tximport (v1.14) and DESeq2 (v1.24) were used to compare gene expression changes between treatments and controls. For putative targets, miR-21-5p targets from TargetScan7.1 and MirDB were downloaded from mirBase and combined into a unified set. onePAGE, as described above, was used to perform gene set enrichment analysis. For empirical targets, we used miR-21 microCLIP binding sites and intersected them with those upregulated upon anti-miR21 transfection.

For smRNA-seq, cutadapt (v3.5) was used to remove linkers and bowtie2 (v2.3.5). UMI-tools (v1.0.0) and UMICollapse were used to account for UMIs and remove duplicates. Annotated small RNAs were then counted and miRNA expression was compared using DESeq2 (v1.24).

Generation of MDA EGFPmiR-21 reporter cell line

The vector backbone for the reporter plasmids was generated from a vector with a bidirectional cytomegalovirus (CMV) promoter-driven lentiviral reporter, expressing eGFP and ΔlnGFR. This vector was a gift from David Erle.[47] The ΔlnGFR ORF in the vector was replaced by a PuroR-T2A-mCherry fusion using Gibson assembly as previously described.[48] In order to generate the eGFP-miR-21 reporter plasmid, two miR-21 binding sites were added to the end of eGFP using the NEBuilder HiFi DNA Assembly Cloning Kit. MDA-MD-231 cells were engineered to stably express the reporter plasmid using lentiviral delivery of the vector.

- hsa-miR-21-5p
- TCAACATCAGTCTGATAAGCTA
- Sequences for reporter cell line generation were obtained from miRBase

FLOW cytometry analysis of miR-21 reporter cell line

MDA EGFPmiR-21 reporter cell lines were seeded at a density of 2×10^4 cells per well of a 96-well plate. Cells were treated for 3 days with serial dilutions of the six most promising drugs (dilutions determined by IC20 curves). Additionally, anti-miR controls (anti-mi21 and non-targeting) were also transiently transfected into cells using Lipofectamine 2000 (Thermo Fisher) according to the manufacturer's protocol. Cells were collected after 3 days and fluorescence output was measured on a BD FACSCelesta flow cytometer.

Animal studies

All animal studies were performed according to IACUC guidelines (IACUC approval number AN194337-01G). Age-matched female NSG mice (Jackson Labs, 005557) were used for metastatic lung colonization assays. These assays were performed with MDA-MB-231 cells, which were seeded at 1.5×10^5 cells in two wells of a six-well plate on day 1. After 24 h 0.1 μM MCULE-9082109585 and 0.5% DMSO were added dropwise to one well each. After 48 h, MCULE-9082109585 and 0.5% DMSO medium was removed and the cells were prepped for tail vein injections. Cells were resuspended in 2 mL of PBS and each mouse received 5×10^4 cells/100 μL of PBS. Metastasis was measured by bioluminescent imaging (IVIS).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100909.

ACKNOWLEDGMENTS

We thank Dr. Kevan Shokat for reviewing the selected molecules and advising on the selection process. H.G. is an Era of Hope Scholar and is supported by the following grants: R01CA240984, R01CA244634, and W81XWH-20-1-0541. H.K. is a Zena Werb Scholar and is supported by T32CA108462. We also thank the UCSF Center for Advanced Technologies, which is supported by UCSF PBBR, RRP IMIA, and NIH1S10OD028511-01 grants.

AUTHOR CONTRIBUTIONS

A.K.A. collected the data, performed the molecular selection, and helped with the implementation of the computational pipeline. M.S. implemented the code and the computational pipeline. H.K. and K.G. performed the in-culture experimental measurements. A.A. analyzed the RNA-seq data. H.G. and J.S.Y. supervised the research. A.K.A., M.S., and H.G. wrote the manuscript.

DECLARATION OF INTERESTS

The work has been patented by the University of Central Florida (UCF) and the University of California San Francisco (UCSF) under the title of "Deep-Learning Based Methods for Virtual Screening of Molecules for Micro Ribonucleic Acid (miRNA) Drug Discovery" (application number: 63/309,132).

Received: February 23, 2023
Revised: November 9, 2023
Accepted: December 7, 2023
Published: January 3, 2024

REFERENCES

1. Peng, Y., and Croce, C.M. (2016). The role of MicroRNAs in human cancer. Signal Transduct. Targeted Ther. *1*, 15004–15009.
2. Kim, J., Yao, F., Xiao, Z., Sun, Y., and Ma, L. (2018). MicroRNAs and metastasis: small RNAs play big roles. Cancer Metastasis Rev. *37*, 5–15.
3. Landskroner-Eiger, S., Moneke, I., and Sessa, W.C. (2013). miRNAs as modulators of angiogenesis. Cold Spring Harb. Perspect. Med. *3*, a006643.
4. Othman, N., and Nagoor, N.H. (2014). The role of microRNAs in the regulation of apoptosis in lung cancer and its application in cancer treatment. BioMed Res. Int. *2014*, 318030.
5. Ma, J., Dong, C., and Ji, C. (2010). MicroRNA and drug resistance. Cancer Gene Ther. *17*, 523–531.
6. Feng, Y.-H., and Tsao, C.-J. (2016). Emerging role of microRNA-21 in cancer. Biomed. Rep. *5*, 395–402.
7. Sicard, F., Gayral, M., Lulka, H., Buscail, L., and Cordelier, P. (2013). Targeting miR-21 for the Therapy of Pancreatic Cancer. Mol. Ther. *21*, 986–994.
8. Zheng, S.R., Guo, G.L., Zhai, Q., Zou, Z.Y., and Zhang, W. (2013). Effects of miR-155 Antisense Oligonucleotide on Breast arcinoma Cell Line MDA-MB-157 and Implanted Tumors. Asian Pac. J. Cancer Prev. APJCP *14*, 2361–2366.

Please cite this article in press as: Keshavarzi Arshadi et al., Functional microRNA-targeting drug discovery by graph-based deep learning, Patterns (2023), https://doi.org/10.1016/j.patter.2023.100909

9. Childs-Disney, J.L., Yang, X., Gibaut, Q.M.R., Tong, Y., Batey, R.T., and Disney, M.D. (2022). Targeting RNA structures with small molecules. Nat. Rev. Drug Discov. *21*, 736–762.
10. Yazdani, K., Jordan, D., Yang, M., Fullenkamp, C.R., Schneekloth, J., Calabrese, D.R., Boer, R.E., Hilimire, T.A., Allen, T.E., and Khan, R.T. (2022). Machine Learning Informs RNA-Binding Chemical Space. Angew Chem. Int. Ed. Engl. *62*, e202211358.
11. Schroeder, S.J. (2018). Challenges and approaches to predicting RNA with multiple functional structures. Rna *24*, 1615–1624.
12. Winkle, M., El-Daly, S.M., Fabbri, M., and Calin, G.A. (2021). Noncoding RNA therapeutics—Challenges and potential solutions. Nat. Rev. Drug Discov. *20*, 629–651.
13. Liu, D., Wan, X., Shan, X., Fan, R., and Zha, W. (2021). Drugging the "undruggable" microRNAs. Cell. Mol. Life Sci. *78*, 1861–1871.
14. Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R.P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Preprint at arXiv.
15. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2023). PubChem 2023 update. Nucleic Acids Res. *51*, D1373–D1380.
16. Sterling, T., and Irwin, J.J. (2015). ZINC 15 – Ligand Discovery for Everyone. J. Chem. Inf. Model. *55*, 2324–2337.
17. (2023). PubChem Bioassay Record for AID 2289: qHTS Assay for Modulators of miRNAs And/or Inhibitors of miR-21, Source: National Center for Advancing Translational Sciences (NCATS), 2010.
18. Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., and Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. *9*, 513–530.
19. (2023). Asinex RNA-Binding - RNA-Targeting Small Molecules.
20. Richard, A.M., Huang, R., Waidyanatha, S., Shinn, P., Collins, B.J., Thillainadarajah, I., Grulke, C.M., Williams, A.J., Lougee, R.R., Judson, R.S., et al. (2021). The Tox21 10K compound library: collaborative chemistry advancing toxicology. Chem. Res. Toxicol. *34*, 189–216.
21. (2023). PubChem Bioassay Record for AID 1347074, Dicer-Mediated Maturation of Pre-microRNA , Source (Center for Chemical Genomics, University of Michigan), 2019.
22. Ryoo, S.-R., Yim, Y.-K., Kim, Y.-K., Park, I.-S., Na, H.-K., Lee, J., Jang, H., Won, C., Hong, S., Kim, S.-Y., et al. (2018). High-throughput chemical screening to discover new modulators of microRNA expression in living cells by using graphene-based biosensor. Sci. Rep. *8*, 11413.
23. Naro, Y., Ankenbruck, N., Thomas, M., Tivon, Y., Connelly, C.M., Gardner, L., and Deiters, A. (2018). Small Molecule Inhibition of MicroRNA miR-21 Rescues Chemosensitivity of Renal-Cell Carcinoma to Topotecan. J. Med. Chem. *61*, 5900–5909.
24. Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. J. Comput. Aided Mol. Des. *30*, 595–608.
25. Liu, K., Sun, X., Jia, L., Ma, J., Xing, H., Wu, J., Gao, H., Sun, Y., Boulnois, F., and Fan, J. (2019). Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. Int. J. Mol. Sci. *20*, 3389.
26. Sakai, M., Nagayasu, K., Shibui, N., Andoh, C., Takayama, K., Shirakawa, H., and Kaneko, S. (2021). Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. Sci. Rep. *11*, 525.
27. Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., and Wang, F. (2020). Graph convolutional networks for computational drug development and discovery. Briefings Bioinf. *21*, 919–935.
28. Chithrananda, S., Grand, G., and Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at arXiv.
29. Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. (2018). Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. Preprint at arXiv.
30. Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Preprint at arXiv.
31. Bai, L., Ong, Y.-S., He, T., and Gupta, A. (2020). Multi-task gradient descent for multi-task learning. Memet. Comput. *12*, 355–369.
32. PubChem Bioassay Record for AID 588342 (2023). qHTS Profiling Assay for Firefly Luciferase Inhibitor/activator Using Purifed Enzyme and Km Concentrations of Substrates (Counterscreen for miR-21 Project), Source: National Center for Advancing Translational Sciences (NCATS).
33. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front. Endocrinol. *9*, 402.
34. Arisan, E.D., Rencuzogullari, O., Cieza-Borrella, C., Miralles Arenas, F., Dwek, M., Lange, S., and Uysal-Onganer, P. (2021). MiR-21 Is Required for the Epithelial–Mesenchymal Transition in MDA-MB-231 Breast Cancer Cells. Int. J. Mol. Sci. *22*, 1557.
35. McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of microRNA targeting efficacy. Science *366*, eaav1741.
36. Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell Rep. *7*, 281–292.
37. Paraskevopoulou, M.D., Karagkouni, D., Vlachos, I.S., Tastsoglou, S., and Hatzigeorgiou, A.G. (2018). microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. Nat. Commun. *9*, 3601.
38. Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. *48*, D127–D131.
39. Goodarzi, H., Elemento, O., and Tavazoie, S. (2009). Revealing global regulatory perturbations across human cancers. Mol. Cell *36*, 900–911.
40. Yang, B., and McJunkin, K. (2020). CRISPR screening strategies for microRNA target identification. FEBS J. *287*, 2914–2922.
41. LumosBio (2023). LumosBio/RiboStrike: Publication Release. v0.0.0 Ed (Zenodo), p. 2023.
42. Torng, W., and Altman, R.B. (2019). Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. J. Chem. Inf. Model. *59*, 4131–4149.
43. Zhao, T., Hu, Y., Valsdottir, L.R., Zang, T., and Peng, J. (2021). Identifying drug–target interactions based on graph convolutional network and deep neural network. Briefings Bioinf. *22*, 2141–2150.
44. Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. (2019). Deep Learning for the Life Sciences (O'Reilly Media).
45. Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., and Massagué, J. (2005). Genes that mediate breast cancer metastasis to lung. Nature *436*, 518–524.
46. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550–621.
47. Amendola, M., Venneri, M.A., Biffi, A., Vigna, E., and Naldini, L. (2005). Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. Nat. Biotechnol. *23*, 108–116.
48. Yu, J., Navickas, A., Asgharian, H., Culbertson, B., Fish, L., Garcia, K., Olegario, J.P., Dermit, M., Dodel, M., Hänisch, B., et al. (2020). RBMS1 Suppresses Colon Cancer Metastasis through Targeted Stabilization of Its mRNA RegulonRBMS1 Modulates RNA Stability to Suppress Colorectal Cancer Metastasis. Cancer Discov. *10*, 1410–1423.

Patterns, Volume *5*

Supplemental information

Functional microRNA-targeting drug discovery by graph-based deep learning

Arash Keshavarzi Arshadi, Milad Salem, Heather Karner, Kristle Garcia, Abolfazl Arab, Jiann Shiun Yuan, and Hani Goodarzi

Supplemental Materials

Molecular Data Preprocessing

After gathering multiple datasets for bioactivity classification, off-target interaction prediction, and inference, the molecules are represented in SMILES format. Each entry is then transformed to be canon with isomeric information included, using the rdkit library in Python (v3.6). Afterwards, the data is cleaned and missing, or duplicate entries are removed. The molecules from the toxicity dataset undergo one more preprocessing step, during which the molecules were desalted, and inorganic molecules were removed. This is since many molecules exist in the toxicity dataset that are salted, are redundant or contain inorganic atoms. Since these molecules are not relevant to this work's screening, they were removed.

Molecular Data Preprocessing: Label Assignment

In order to assign bioactivity labels to each molecule from the PubChem datasets, the two columns of "outcome activity" and "phenotype activity" were examined. In most assays, the datapoints with "Activator" phenotype were removed, since inhibitors alone were the desirable positive label for this work. In two cases (AIDs 504466 and 624202) the inhibitors were removed instead, since inhibitor and activators had the opposite meaning compared to other assays. For the remaining molecules, the outcome activity column is used to assign bioactivity labels, with 'Active' and 'Inconclusive' outcomes being labeled as 1, and 'Inactive' and 'Unspecified' outcomes being labeled as 0.

Molecular Data Preprocessing: Data Splitting

The combined dataset, toxicity dataset, and DICER dataset are all split into training (80%), validation (10%), and test sets (10%) using DeepChem, based on the molecular scaffolds[1]. Splitting based on the scaffolds allows a larger distinction to exist between the molecules of each split, resulting in a more practical model for real-world applications where inference data is often from a different distribution than the training data. Moreover, splitting the combined dataset instead of each of the virtual screening datasets allows for the test dataset to remain constant throughout all types of training trials, resulting in a fair comparison between the models.

Hyper-Parameter Optimization: Training

After the data is gathered and preprocessed to have binary labels and canonical SMILES, hyper-parameter optimization is performed to find the suitable model architecture for each training task. Using the validation set, different model architectures are evaluated for their performance. In a grid search manner, the number of graph convolutional layers, the size of each layer, the size of the feed-forward layer, the batch size, and the learning rate are altered and the best performing set of parameters on the validation set is taken to initiate the training. This hyper-parameter optimization is extended to the training epoch selection, where the epoch with the best performance on the validation set is chosen as the final model.

Hyper-Parameter Optimization: Results

The results are shown in Table S1 for the virtual screening models. All models are trained for 150 epochs and the best epoch is identified using the AP score on the validation set.

As seen in Table S1, the hyper-parameter search arrives at similar optimum architectures for all datasets, with the model for the recommended task having a smaller batch size and training duration. This is intuitive since this training scenario has fewer tasks and a smaller training dataset.

Table S1- Optimum hyper-parameters for each training scenario of the virtual screening models. Best epochs are found through early stopping monitoring the validation set.

Model	Convolution Layers	Dropout	Learning Rate	Batch Size	Dense Layer Size	Best Epoch
Single Task	512, 512, 512	0.1	0.0001	128	1024	110
Multitask - Cancer	512, 512, 512	0.1	0.0001	128	1024	139

Multitask - All	512, 512, 512	0.1	0.0001	128	1024	145
Recommended Tasks	512, 512, 512	0	0.0001	32	1024	34

The best performing model was the result of multitask learning on the recommended tasks, which will be explored in the next section. The Receiver-Operator Curve (ROC) and Precision Recall Curve are shown in Figure S1.

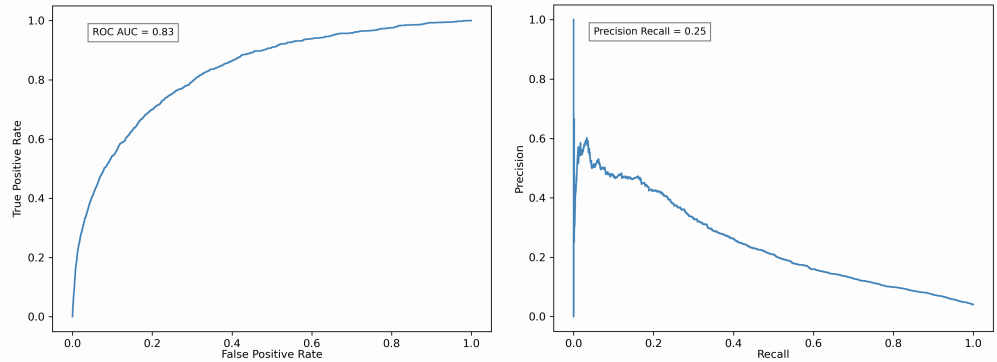


Figure S1: Receiver-Operator Curve (left) and Precision Recall Curve (Right) of the best performing model. This model was trained using the tasks recommended by the task recommender algorithm. The high precision section of the precision-recall curve is where the model operates to select candidates with low uncertainty during virtual screening.

Task Recommendation Results: Tailored Tasks for miR-21

The task recommendation algorithm is implemented to find tasks within the input dataset that when used for data augmentation, would result in a better performance for our target task of miR-21. Table S2 lists the tasks that were chosen and recommended.

Table S2 – The recommended and optimized tasks for multitask learning for miR-21 virtual screening. Each assay is represented by its Assay Identifier (AID) in PubChem.

Recommended Task Name	Importance
AID-1458: qHTS Assay for Enhancers of SMN2 Splice Variant Expression	Other/Unrelated
AID-485297: qHTS Assay for Rab9 Promoter Activators	Other/Unrelated
AID-485313: qHTS Assay for NPC1 Promoter Activators	Other/Unrelated
AID-588342: qHTS profiling assay for firefly luciferase inhibitor/activator using purified enzyme and Km concentrations of substrates (counterscreen for miR-21 project)	Counter Screen
AID-504466: qHTS screen for small molecules that induce genotoxicity in human embryonic kidney (HEK293T) cells expressing luciferase-tagged ELG1	Other/Unrelated
AID-624202: qHTS Assay to Identify Small Molecule Activators of BRCA1 Expression	Other/Unrelated
AID-2289: qHTS Assay for Modulators of miRNAs and/or Inhibitors of miR-21	Main Screen

Toxicity and DICER Inhibition Modeling Results

For the case of DICER inhibition prediction, a single task model is trained on the available data. On the other hand, the toxicity dataset contains multiple assays, which requires the same training pipeline (multitask learning as well as task recommendation) as the miR-21 virtual screening model to be reimplemented. In this process after a model is trained on all of the toxicity tasks, the prediction-based recommendation algorithm returns 5 tasks that would assist the performance of the model on the HepG2 target task.

Comparison to ChemBERTa: Language Modeling for Small Molecules

RiboStrike relies on GCNN architecture for representation of compounds and modeling due to its simplicity, scalability, and performance. In this section, the GCNN architecture performance is compared to two state-of-the-art approaches, fingerprint based and language model based classification of compounds.

In recent years, language models have gained popularity in modeling text within the natural language processing field as well as other sequence-based data such as proteins and even small molecules within the biology and chemistry fields. These models have the capability to learn from unlabeled sequences solely by relying on a mechanism called "masked token prediction", where certain words or "tokens" (e.g. amino acids, nucleotide bases, or atoms within a SMILES string) are omitted and the model is asked to predict the missing tokens. By doing so, the model can learn the grammar and syntax of the sequences, and with this knowledge, create meaningful representations for these sequences. ChemBERTa[2] is a language model trained on 77 million SMILES strings from PubChem, which deploys masked token prediction and has a distinctive feature space to distinguish small molecules given their SMILES.

To assess the performance of our GCNN-based model, we compare it to ChemBERTa in two manners, feature extraction and finetuning. During feature extraction, the outputs of the last layer of the model were taken as features then classified with a Random Forest model. During fine tuning, the pre-trained model was taken from HuggingFace repository and modified to be fit for a classification task by adding one fully connected layer to the end of the model (with 32 neurons). The weights of the base model were frozen during training, except for the Layer Norm parameters to perform selective finetuning[3]. The model was trained for 20 epochs to predict the binary labels of the miR-21 task. The results are shown in Table S2. The model is also compared to the traditional method of Extended Connectivity Fingerprint (ECFP4) classification, which relies on encoding the existence of subgraphs of the compound in a binary vector.

Table S2 – Performance of this work's best model compared to ChemBERTa on the test set of the miR-21 dataset. Models are trained and tested on the same datasets.

Model	Method	ROC-AUC	Accuracy	Recall	Precision	AP
RiboStrike	Multitask learning	0.8305	90.39	49.75	20.93	**24.72**
ChemBERTa - Fine Tuned	Masked language modeling	0.7547	88.57	36.40	14.22	11.89
ChemBERTa - Features	Masked language modeling	0.5010	95.83	0.35	8.93	4.05
ECFP4	Fingerprint	0.5039	95.94	0.85	38.71	4.33

As it can be seen from Table S2, the approach proposed in this work can empirically outperform ChemBERTa in both modes and ECFP-based classification, with the help of multitask learning as well as task recommendation. One reason for ChemBERTa's lower performance can be the resolution of the fingerprints it offers. As we have seen from Figure 4B, the current dataset has massive similarity between the positive and negative ground truth data points. This similarity would necessitate high-resolution fingerprints to be leveraged to discriminate between the data points of different classes, while ChemBERTa's inner representations are general and lack this resolution. Overall, the comparison made here is for using ChemBERTa as a feature extractor as it is the prevalent approach, it is noteworthy to mention that the pipeline of RiboStrike can be model-agnostic and the principals of multitask learning and task recommendation can be extended to any model, even pre-trained language models such as ChemBERTa.

Overlap Between Positive and Negative Training Data: Improvements with RiboStrike

As previously depicted in Figure 4, the positive and the negative data clusters in the miR-21 training dataset have a large area of overlap in their UMAP projection. This overlap can decrease the model's performance and a model with a more disjoint feature space should have more distinguishing representations and intuitively have better performance. In this section, this overlap is measured and used to compare the representation from RiboStrike, ChemBERTa, and ECFP in their ability to divide the positive and negative data points. To measure this overlap, we use the Silhouette Coefficient, commonly used in clustering techniques to measure how well each data point fits its assigned cluster via measuring the inter-cluster cohesion and intra-cluster separation. The representation space as well as the measured Silhouette Coefficients are reported in Figure S2.

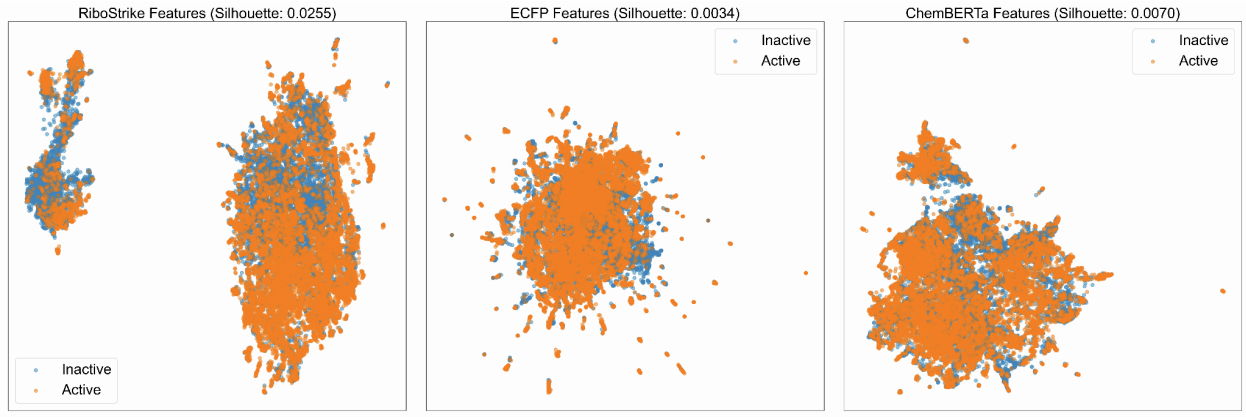


Figure S2: UMAP of representations from different models and feature extraction techniques. The positive and negative instances in the miR-21 training data are first featurized, then depicted using UMAP to visualize the overlap between these classes. Silhouette Coefficient is measured to compare this overlap empirically (higher has less overlap).

The results indicate that overlap exists between all three methods due to the nature of this classification task. However, RiboStrike has a higher Silhouette Coefficient and performs better at distinguishing between the active and inactive compounds.

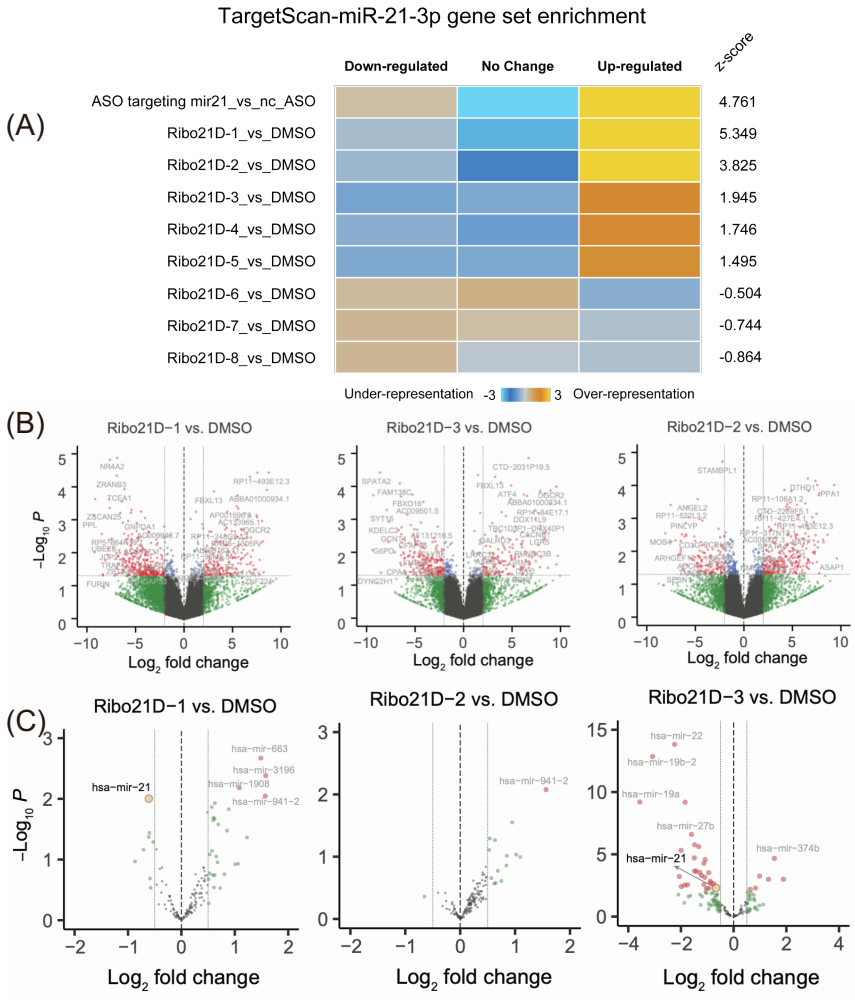
Molecular IDs

The name, IDs, and the SMILES for the molecules selected in this work for experimental validation are mentioned in Table S3.

Table S3 – The selected molecules by the RiboStrike pipeline. Molecules within the text are referred to by their assigned names and can be found through their molecular IDs.

Molecular ID	Name	Smiles
MCULE-9082109585_vs_DMSO	Ribo21D-1	Cc1cc(C(=O)CSc2ccccn2)c(C)n1C
MCULE-7511323413_vs_DMSO	Ribo21D-2	O=C(NCCc1ccc(F)cc1)NC12CC3CC(CC(C3)C1)C2
MCULE-7127547568_vs_DMSO	Ribo21D-3	Cc1cc2nc(N3CCN(C)CC3)sc2cc1C
MCULE-	Ribo21D-4	COc1cc(C)ccc1CO/N=C(\N)c1nonc1N

7903461821_vs_DMSO		
AEM14465585_vs_DMSO	Ribo21D-5	CC1=CC=C(O1)C1=C(C=NC(NCCCN2C=CN=C2)=N1)C1=CC(C)=NO1
CSMS00103425726_vs_DMSO	Ribo21D-6	O=C(NC[C@H]1CCCO1)c1cc(C2CC2)on1
AEM14465594_vs_DMSO	Ribo21D-7	CN(C)CCCNC1=NC(C2=CC=C(C)O2)=C(C=N1)C1=CC(C)=NO1
MCULE-1056460601_vs_DMSO	Ribo21D-8	CC1CCN(C(=O)CCSc2ccccc2)CC1



Supplementary Figure S3. Experimental Validations for the top compounds. (A) gene-set enrichment analysis to sort candidate molecules based on upregulating effect on miR-21 target genes in our QuantseqPOOL results. QuantseqPOOL is a multiplexed 3'-end RNA sequencing assay that provides a scalable approach for

performing gene expression quantification for many samples simultaneously. Using this approach, we used RNA-seq statistical models to cluster genes into three groups in each treatment: those that are down-regulated, are up-regulated, or show no significant and substantial change in expression. We have then presented the enrichment of miR-21 targets in each group and reported the z-score associated with this gene-set enrichment analysis[4]. The first row shows the results for the anti-miR21 ASO relative to a non-targeting ASO control. Other rows depict the comparison of cells treated at IC20 with each drug relative to DMSO control in MDA-MB-231 cells. We selected the top three drugs for additional study using high coverage RNA-seq and small RNA-seq. **(B)** Volcano plots for RNA seq analysis of the top three compounds. Samples were profiled in biological triplicates and compared to DMSO control. The x-axis shows log2 fold-changes and the y-axis shows the -log10 of P. **(C)** Volcano plots comparing miRNA profiles of treated cells to DMSO control. From the same total RNA in B, we also generated small RNA sequencing libraries that we used to quantify changes in miRNAs, especially miR-21 itself. Ribo21D-1 and -3 resulted in a significant decrease in miR-21 levels; in case of Ribo21D-3 this effect was not specific as a large number of miRNAs are simultaneously downregulated. Ribo21D-2 did not show a significant reduction in any miRNAs.

miR-122 Interaction Prediction

Similar to interaction with Dicer, off-target interactions with other miRNAs (e.g. miR-122) are also not desired. The RiboStrike pipeline was used to train a single-task inhibition predictor of miR-122. The confusion matrix of the miR-122 model as well the training and validation metrics during training are shown in Figure S4.

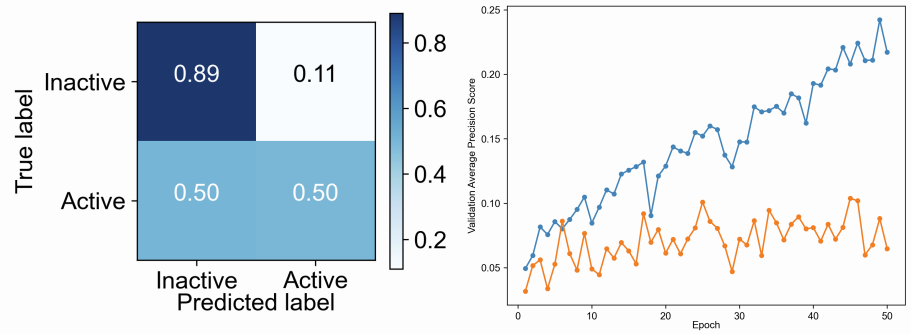


Figure S4 – Results of the auxiliary counter-screen model trained on miR-122 molecular data. (Left) The confusion matrix of the trained model); **(Right)** The monitored training (blue) and validation (orange) average precision score during training, used for early stopping of the model.

This auxiliary model is used as a virtual counter-screen against miR-122, assisting in the specificity assessment of the main model. The results of these predictions are reported in Table S4. These results show that 7 out of the top 8 compounds predicted by the pipeline are predicted to not interact with miR-122, with none of the top three candidates failing the virtual counter-screen.

Table S4 –Predicted interaction of the selected compounds with miR-122 and the uncertainty in the prediction. No predicted interaction with low uncertainty is better as off-target interaction specially with other miRNAs are not desired.

Name	Prediction	Uncertainty
Ribo21D-1	0	0.68
Ribo21D-2	0	0.42
Ribo21D-3	0	0.47
Ribo21D-4	0	0.29

Ribo21D-5	1	0.41
Ribo21D-6	0	0.32
Ribo21D-7	0	0.29
Ribo21D-8	0	0.29

Task Recommendation Helps Narrow the Scope of Multitask Learning

In RiboStrike, multitask learning is used with the intent to share the representation learning process among multiple datasets, and learn representations that are distinctive for classification yet generalizable among tasks, increasing the performance of the resulting model. However, in practice, there exist cases where multitask learning does not necessarily result in improved performance for all involved tasks. E.g. when tasks have opposing gradients during training. To that end, RiboStrike uses task recommendation to narrow the scope of multitask learning and to recommend datasets that when combined result in improved performance for a target task. Firstly, a multitask model is trained on all available datasets, leveraging graph convolution layers to extract representations from the input graphs and create neural fingerprints as shared representations. Once trained, the output of the multitask model is used for ranking the input datasets in regards to their importance from the view of the task recommendation algorithm. The top ranking datasets are gathered and grouped to form a new multitask dataset, and the training is performed once more. The final model includes tasks that aid the target task in regards to its performance, and is used alongside other filtering models, uncertainty values, and clustering to select drug-like candidates for experimental validation. Figure S5 illustrates the stages of multitask learning and task recommendation in the RiboStrike pipeline.

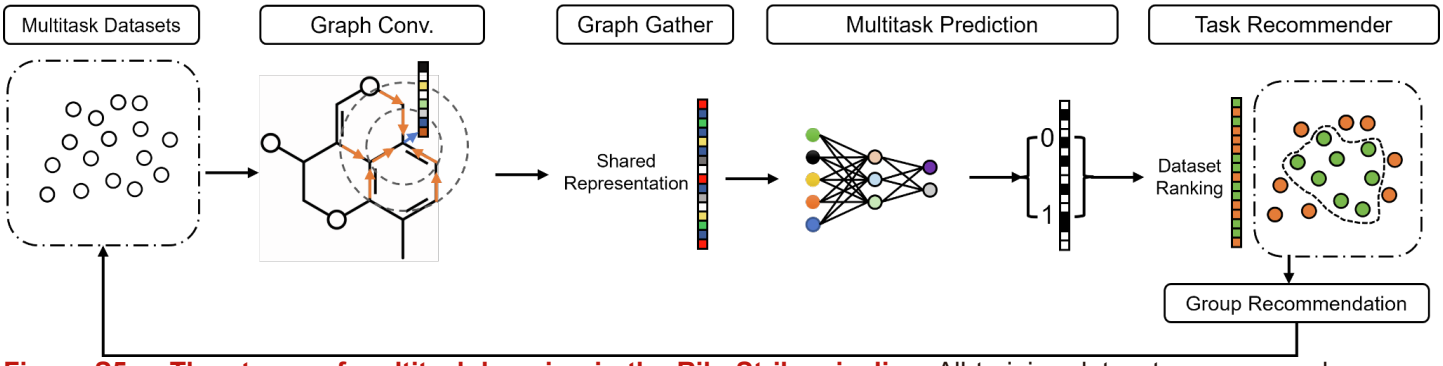


Figure S5 – The stages of multitask learning in the RiboStrike pipeline. All training datasets are grouped initially and used to train a graph convolutional model and learn shared representations, with the intent of gaining generalization in the representations. The trained multitask model is used to make predictions on the target task, allowing the task recommender algorithm to rank each sub-model and its prior dataset. The top-ranking tasks are grouped to form a new multitask dataset to train the final model.

Supplemental References

[1] G. W. Bemis and M. A. Murcko, "The Properties of Known Drugs. 1. Molecular Frameworks," *Journal of Medicinal Chemistry,* vol. 39, no. 15, pp. 2887-2893, 1996/01/01 1996, doi: 10.1021/jm9602928.

[2] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885,* 2020.

[3] M. Salem, A. Keshavarzi Arshadi, and J. S. Yuan, "AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning," *BMC bioinformatics,* vol. 23, no. 1, pp. 1-17, 2022.

[4] Hani. G, Olivier. Elemento, Saeed. T,Revealing Global Regulatory Perturbations across Human Cancers, DOI 10.1016/j.molcel.2009.11.016