Journal of Cheminformatics

RESEARCH Open Access

DeepRNA-DTI: a deep learning approach for RNA-compound interaction prediction with binding site interpretability

Haelee Bae[1] and Hojung Nam[1,2,3*]

Abstract

RNA-targeted therapeutics represent a promising frontier for expanding the druggable genome beyond conventional protein targets. However, computational prediction of RNA-compound interactions remains challenging due to limited experimental data and the inherent complexity of RNA structures. Here, we present DeepRNA-DTI, a novel sequence-based deep learning approach for RNA-compound interaction prediction with binding site interpretability. Our model leverages transfer learning from pretrained embeddings, RNA-FM for RNA sequences and Mole-BERT for compounds, and employs a multitask learning framework that simultaneously predicts both presence of interactions and nucleotide-level binding sites. This dual prediction strategy provides mechanistic insights into RNA-compound recognition patterns. Trained on a comprehensive dataset integrating resources from the Protein Data Bank and literature sources, DeepRNA-DTI demonstrates superior performance compared to existing methods. The model shows consistent effectiveness across diverse RNA subtypes, highlighting its robust generalization capabilities. Application to high-throughput virtual screening of over 48 million compounds against oncogenic pre-miR-21 successfully identified known binders and novel chemical scaffolds with RNA-specific physicochemical properties. By combining sequence-based predictions with binding site interpretability, DeepRNA-DTI advances our ability to identify promising RNA-targeting compounds and offers new opportunities for RNA-directed drug discovery. The codes and data are publicly available at https://github.com/GIST-CSBL/DeepRNA-DTI/.

Scientific contribution

We present DeepRNA-DTI, a novel sequence-based deep learning framework that simultaneously predicts RNA-compound interactions and identifies nucleotide-level binding sites through multitask learning, providing mechanistic interpretability unlike existing methods that focus solely on interaction prediction. By integrating comprehensive datasets from PDB and literature sources with pretrained embeddings, our model demonstrates superior generalization across diverse RNA subtypes without requiring subtype-specific training. Large-scale virtual screening against oncogenic pre-miR-21 successfully identifies both known binders and novel RNA-specific chemical scaffolds, demonstrating practical utility for RNA-targeted drug discovery.

Keywords RNA compound interaction, Drug target interaction, Binding site, Deep learning, Multi task learning

*Correspondence:
Hojung Nam
hjnam@gist.ac.kr
Full list of author information is available at the end of the article

Introduction

The therapeutic landscape in modern drug discovery remains predominantly focused on protein targets, which are encoded by merely 2% of the human genome. Of these proteins, a mere 15% are disease-related and possess druggable binding pockets. As a result, less than 0.02% of all genes are currently targetable with conventional therapeutics. To address a wider range of diseases, expanding the target space beyond proteins is crucial [1–7].

RNA molecules, encoded by approximately 70% of all genes, play vital roles in regulating various biological processes [8–10]. Given the limitations of protein-targeted therapies, RNA presents a promising alternative for therapeutic intervention, as it is encoded by a significantly larger portion of the genome. Additionally, RNA's ubiquity across a wide range of organisms, including mammals, viruses, bacteria, and fungi, further enhances its potential as a therapeutic target [11]. Moreover, some drugs designed to target proteins may unintentionally interact with RNA, causing off-target effects. Therefore, it is crucial to evaluate potential RNA off-target interactions to improve drug safety and efficacy [12].

RNA can be targeted using small molecules or oligonucleotides, each with distinct advantages and limitations. Small molecules, due to their low molecular weight, facilitate efficient cellular uptake and delivery with minimal immunogenicity, but they rely on the presence of accessible RNA binding pockets and may encounter challenges in achieving high selectivity. Oligonucleotides, while offering high selectivity through complementary base pairing, face significant obstacles in cellular uptake because of their larger size. Consequently, small molecules are often preferred for RNA targeting when suitable binding pockets are available [13–15]. Recent advancements in deep learning have further enhanced the ability to predict RNA binding sites, improving the feasibility of small-molecule-based RNA targeting strategies [16–26].

Several studies have developed computational models to predict RNA-small molecule interactions. Early approaches utilized RNA structural information for ligand recommendation. Inforna2.0 [27], R-Bind [28] and RNALigands [29] use RNA secondary structures as input to recommend ligands based on RNA similarity, assisting in ligand selection. RNAmigos [30] employs graph representations of RNA binding sites and generates MACCS fingerprints, offering insights into the properties of compounds that bind to specific sites, though it does not provide exact structural details.

More recent developments have focused on quantitative interaction prediction. Cai et al. [31] developed a quantitative structure activity relationship model to predict affinity using compound features, but is specific to the HIV-1 TAR RNA, limiting its generalizability to other RNAs. ROBIN [32] offers experimental interaction data for various RNAs, DNAs, and compounds. They make a machine-learning model to predict whether compounds bind RNA or proteins. RSAPred [33] predicts affinity as pKd values using RNA and compound features with a linear regression model and is the first general model applicable to unseen RNAs, though it constructs separate models for each RNA subtype.

Structure-based approaches have also shown promise but require 2D or 3D structural information. RLAffinity [34] predicts affinity using 3D RNA-compound complexes but requires available 3D structures, or otherwise docking simulations are needed. DeepRSMA [35] predicts affinity as pKd values, encoding RNA using predicted secondary structure based graph embeddings and sequence based embeddings, while compounds are represented with graph and sequence-based embeddings. ZHMol-RLinter [36] employs double layer stacking of random forest models using RNA sequence, structural, geometric, physicochemical features and compound fingerprint. The model predicts small molecule-nucleotide interaction and small molecule-motif interaction. RNAmigos2 [37] predicts binding score from RNA binding pocket graphs and compound features. GerNA-Bind [38] leverages 1D, 2D and 3D RNA embeddings along with 2D and 3D compound embeddings for binary classification, while also predicting binding sites via pairwise contact matrices and incorporating evidential learning for uncertainty estimation.

More recently, several advanced sequence-based methods have emerged to address RNA-compound interaction prediction. BioLLMNet [39] employs foundation models for RNAs and compounds to predict binding affinity. EMMPTNet [40] encodes RNA using k-mer distribution and de Brujin graph, compounds using molecular descriptors and atomic level features, and predicts affinity. RNAsmol [41] use multiview CNN for RNA, Graph diffusion convolution for compounds, combined with attention for classification. To address the data scarcity, they propose data perturbation and augmentation strategies.

In summary, existing models are either restricted in their application to novel RNAs and compounds or depend on known 3D structural information. Additionally, sequence-based models often lack interpretability, limiting their utility in elucidating RNA-compound interactions.

To address these limitations, we propose DeepRNA-DTI, the generalized deep learning model for sequence-based RNA-compound interaction prediction (Fig. 1). Considering the inherent flexibility and conformational dynamics of RNA molecules [42–44], sequence-based

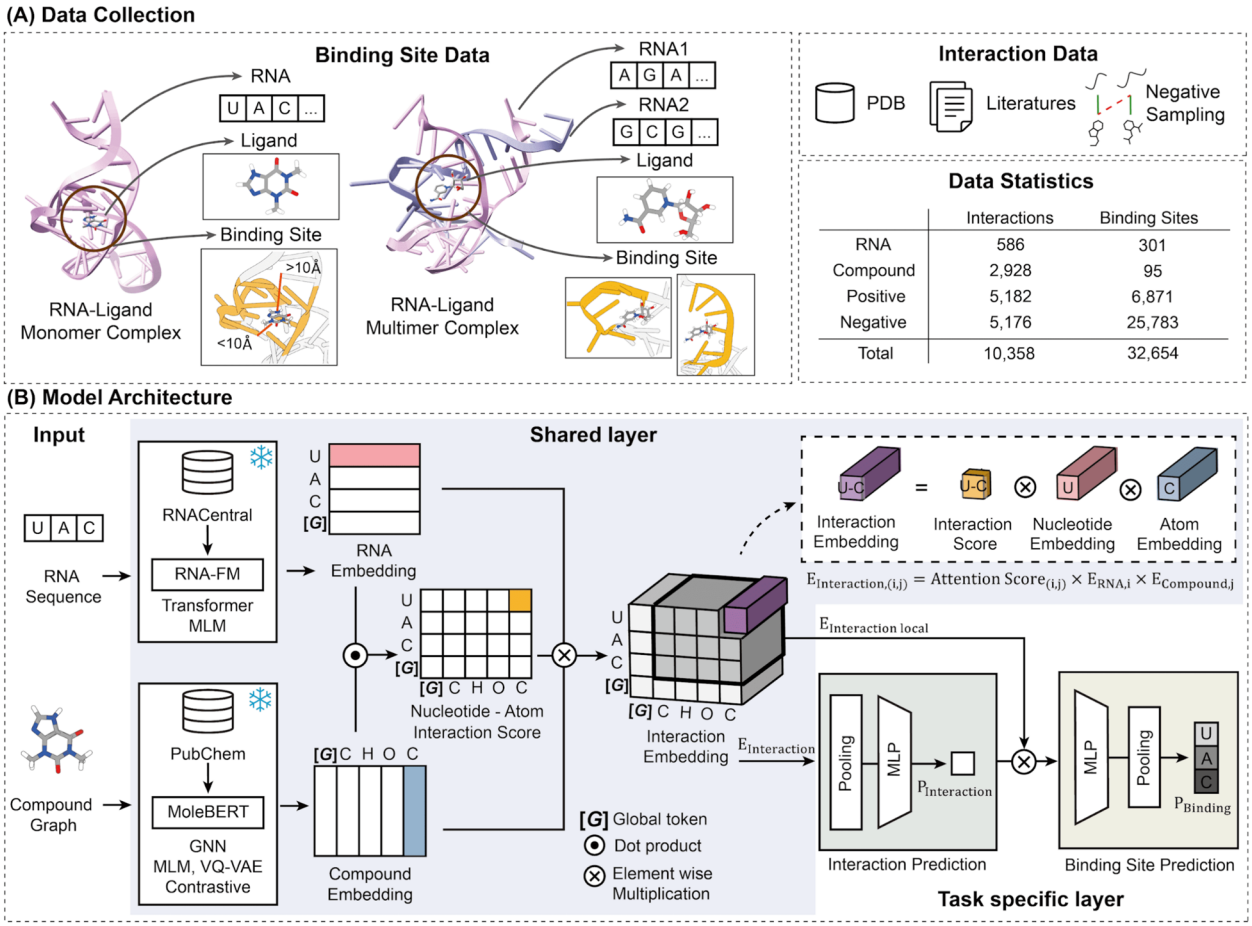

Fig. 1 DeepRNA-DTI Overview. **A** Data Collection: Interaction data were gathered from the Protein Data Bank (PDB) and literature sources. Binding site data were obtained by identifying RNA sequences within 10 Å of ligands for both monomeric and multimeric RNA structures. **B** Model Architecture: RNA sequences and compound structures are encoded using RNA-FM and Mole-BERT, respectively. Shared layers (blue box) generate interaction embeddings, while task-specific layers (purple, pink box) predict interactions and binding sites simultaneously

models may offer an enhanced ability to represent RNA characteristics, where structure-based models might struggle to capture such variability. This also makes the model applicable to RNAs without structures. Our goal is to predict drug-target interactions (DTI) and binding sites simultaneously via multitask learning to enhance both accuracy and interpretability. We incorporate Protein Data Bank (PDB) data to refine binding site predictions, offering more precise and interpretable insights into RNA-compound interactions. This enables DeepRNA-DTI to leverage the flexibility of sequence-based prediction while integrating structural insights, ultimately improving both the accuracy and interpretability of DTI predictions. This comprehensive approach can expand the potential interaction space and offers valuable contributions to RNA-targeted drug discovery.

In the following sections, we describe dataset construction, model architecture, and training strategy, followed by performance evaluations on interaction and binding site prediction tasks, and a large scale screening case study on pre-miR-21.

Materials and methods

Dataset collection and processing

We constructed our dataset by combining data from the PDB and literature sources [31–33, 45] to create a comprehensive collection of RNA-compound interactions. The dataset encompasses both interaction data and binding site information, with careful processing applied to ensure data quality and relevance. Table 1 summarizes the composition of the final dataset. The columns indicate the prediction task, dataset source and RNA oligomer type, number of unique RNAs and compounds, and the number of positive and negative interaction pairs for interaction prediction or residues for binding site prediction.

We began by querying the PDB database for structures containing both RNA and ligand components. Using the

Table 1 Dataset statistics

Task[1]	Dataset[2]		RNA[3]	Compound[4]	Positive[5]	Negative[6]
Interaction	PDB	Monomer	154	62	234	233
		Dimer	105	46	128	125
		Trimer	17	10	22	22
		Tetramer	29	11	33	33
	ROBIN [32]	Monomer	25	1967	3386	3390
		Dimer	2	200	200	200
	RSAPred [33]		271	690	1017	1039
	fingeRNAt-ml [45]		6	161	122	127
	Cai et al. [31]		1	47	40	7
	Total		586	2928	5182	5176
Binding Site	PDB		301	95	6871	25,783

[1] Types of the prediction task

[2] Types of Dataset source and RNA oligomer

[3] Number of unique RNAs

[4] Number of unique compounds

[5] Number of positive pairs

[6] Number of negative pairs

RCSB API, we retrieved the corresponding PDB files, excluding entries that were inaccessible. The structures were categorized based on their oligomer count as either monomers or multimers. We retained only RNA chains, removing all non-RNA components, and excluded structures containing non-canonical sequences ('X' or 'N') or RNA sequences exceeding 200 nucleotides in length. When multiple structures contained identical RNA-ligand pairs, we retained only the one with the highest resolution. For the remaining structures, we downloaded assembly files from the RCSB. When multiple assemblies or models were available for an identical structure, we selected the first instance. To determine RNA-ligand interactions, we calculated atomic distances between RNA chains and ligands within each structure. An interaction was classified as positive if any atoms between an RNA chain and a ligand were within 10 Å of each other. We excluded compounds that could not be parsed using RDKit from the dataset. For the Drug-Target Interaction (DTI) data, we generated balanced datasets by creating negative interaction samples. These were constructed by randomly pairing compounds from positive interactions with RNAs from the positive set, ensuring the number of negative samples matched the positive interactions. The binding site data utilized only positive interaction data from PDB, as these contained explicit binding site annotations based on structural proximity. Binding sites were defined as nucleotide residues where any atom was within 10 Å of any ligand atom, while nucleotide residues located more than 10 Å away from the ligand atom were considered negative binding sites.

From the ROBIN dataset [32], we included only explicitly specified RNA entries and retained only canonical RNA sequences. We excluded compounds without RNA interactions and generated balanced negative interactions through random sampling. The RSApred dataset [33] provided DTI data, where we labeled interactions as positive for $pKd > 4$ and negative for $pKd \leq 4$. The pKd value is the dissociation constant (Kd) transformed into log space. A detailed explanation and formula are provided in Supplementary 1.1. We generated additional negative pairs through random RNA-compound pairing to achieve balance. The fingeRNAt-ml dataset [45] included six RNA targets with both positive and negative interactions, from which we used only 'real' type interactions and retrieved RNA sequences from PDB using their API. For the Cai et al. dataset [31], we labeled interactions as positive for $pKd > 4$ and negative for $pKd \leq 4$.

The final combined dataset represents a diverse collection of RNA-compound interactions, comprising 586 unique RNA sequences and 2928 unique compounds. The dataset contains 5182 positive interactions and 5176 negative interactions, along with 6871 positive binding sites and 25,783 negative binding sites. The distribution of RNA type and compounds across different RNA types are provided in Supplementary 1.2 and 1.3.

Data splitting for training, validation, and testing

To ensure the rigorous evaluation of model performance, we designed four distinct data split schemes, Unseen Pair, Unseen RNA, Unseen Compound, and Unseen Both.

A separate model was trained for each setting. While it is theoretically possible to train a single universal model and test it across all unseen settings, this approach was not feasible due to the limited number of available RNA-compound interaction samples. Combining all unseen scenarios into a single training process would reduce the number of examples, resulting in unstable training and unreliable evaluation. Therefore, we adopted a scenario-specific training strategy to ensure fair and robust comparison.

In all settings, the dataset was first divided into training and test sets, and five-fold cross validation was performed on the training set. Five models were evaluated on the test set, and the final performance was averaged.

To establish in-distribution evaluation, we constructed an unseen pair split, where RNA or compound may appear in both the training and test set, but not in the same pair combination. Given the scarcity of monomer structures in the PDB dataset, we utilized multimer structures as a form of data augmentation to enhance learning from a broader range of RNA sequences. For the PDB dataset, we divided the monomer data into six-folds, using five folds for cross-validation during training and reserving the remaining fold for testing. The multimer data were split into five folds and incorporated solely into the training process to augment the learning without affecting test evaluation. We applied a similar strategy to the ROBIN dataset [32], where monomer data were partitioned into six-folds, five for cross-validation during training and one for testing. The multimer data from this dataset were divided into five folds and used exclusively for training purposes. For the dataset from RSApred [33] and fingeRNAt-ml [45], which also contain only monomer structures, we partitioned into six folds – five for cross validation during training and one for testing. For the dataset from Cai et al. [31], which consisted entirely of monomer structures, we maintained the original train/test split provided in their publication to ensure comparability with previous research. The samples designated as test data in the original study were preserved for testing in our work, while the training data were used accordingly. To prevent bias from sequence length differences, RNA sequences were distributed across folds to ensure similar length distribution (Supplementary 1.4).

To evaluate the out-of-distribution (OOD), the data were split based on entity level similarity. For Unseen RNA setting, RNA sequences were clustered using CD-HIT at 80% identity, and train and test sets derived from distinct clusters. For Unseen Compound setting, Compounds were clustered using Butina clustering on ECFP fingerprints with a tanimoto cutoff of 0.4. Train and test sets were drawn from distinct compound clusters. For Unseen Both setting, both RNA and compound clusters were simultaneously excluded from training, ensuring that neither appeared in the training data. In all OOD settings, the binding site and DTI datasets were aligned to maintain consistent entity assignments. RNA or compound clusters assigned to training data were identically assigned in the DTI dataset. All multimer data were included in the training set to enhance learning. Remaining data were distributed by count into train and test sets.

Model architecture

We propose DeepRNA-DTI, a deep learning model designed to predict RNA-compound interactions using RNA sequences and compound structures as inputs (Fig. 1). The architecture employs a multitask learning approach that simultaneously optimizes the prediction of binding sites and drug-target interactions (DTI), where binding site prediction serves as an auxiliary task that enhances DTI prediction performance while providing nucleotide-level interpretability.

Overview

DeepRNA-DTI consists of shared layers and task-specific layers. The shared layers encode a common pattern of binding between RNA sequences and compound graphs via attention mechanisms, capturing complex relationships between the two modalities. The task-specific layers comprise (1) a binding site prediction layer and (2) an interaction prediction layer, each dedicated to learning patterns specific to their respective tasks.

This multitask learning enables the model to capture complementary information from both tasks, potentially leading to a more comprehensive understanding of RNA-compound interactions. Additionally, predicting binding sites alongside interactions can enhance the interpretability of the model's predictions, providing valuable insights even when the RNA structure is unknown.

Input representation

RNA sequences are encoded using RNA-FM [46], a pre-trained Transformer model trained on approximately 23 million non-coding RNA sequences from RNACentral [47] using masked language modeling (MLM) tasks. It has demonstrated excellent performance in various tasks, including 3D structure prediction. To leverage the pre-trained embeddings effectively, all layers of RNA-FM are frozen during training.

For an RNA sequence of length L_{RNA}, the output embedding has a shape of (L_{RNA}, dim).

These RNA embeddings are passed through an RNA-specific linear layer followed by a ReLU activation function, reducing the dimensionality to $(L_{RNA}, 128)$.

Compound graphs are encoded using Mole-BERT [48], a model trained on 2 million compounds from

ZINC using a Vector Quantized Variational Autoencoder (VQ-VAE) with MLM and contrastive learning tasks. All layers of Mole-BERT are frozen to utilize the pretrained embeddings. Using the compound graph as input, the output embedding has a shape of (N_{atom}, dim), where N_{atom} represents the number of atoms in the compound. The compound embeddings are processed through a compound-specific linear layer and ReLU activation, resulting in embeddings of shape $(N_{atom}, 128)$. The detailed rationale for the selection of RNA-FM and Mole-BERT is described in Supplementary 2.1.

Interaction embedding

To incorporate global information, the average RNA embedding and the average compound embedding are computed and concatenated to each respective RNA and compound embedding. This results in RNA representations of shape $(L_{RNA} + 1, 128)$ and compound representations of shape $(N_{atom} + 1, 128)$. The interaction between RNA and compound is modeled by computing attention scores using the following formula:

$$AttentionScore = E_{RNA} W_1 W_2 E_{Compound}^T$$

where E_{RNA} and $E_{Compound}$ are the RNA and compound embeddings, respectively, and W_1 and W_2 are learnable weight matrices.

The interaction embeddings $E_{Interaction}$ of shape $(L_{RNA} + 1, N_{atom} + 1, 128)$ are obtained by:

$$E_{Interaction,(i,j)} = AttentionScore_{(i,j)} \times E_{RNA,i} \times E_{Compound,j}$$

where i indexes the RNA sequence positions and j indexes the compound atoms. This element-wise multiplication creates embeddings representing the interaction strength between each RNA nucleotide and each compound atom.

These interaction embeddings are subsequently processed through two linear layers followed by a ReLU activation function to produce the final representation used for downstream tasks.

Interaction and binding site prediction layer

The DeepRNA-DTI model employs two specialized layers to perform distinct prediction tasks: interaction prediction and binding site prediction.

For interaction prediction, the model derives a global representation of the RNA-compound interaction by averaging the interaction embedding across all RNA nucleotides and compound atoms. This process yields E_{Pool} with a shape of (1,32), calculated as:

$$E_{Pool} = \frac{1}{L_{RNA} \times N_{atom}} \sum_{i=1}^{L_{RNA}} \sum_{j=1}^{N_{atom}} (E_{Interaction})$$

This global embedding is then processed through a neural network consisting of two linear layers with a ReLU activation function in between, resulting in a single scalar value representing the probability of interaction:

$$P_{Interaction} = \sigma(Linear(ReLU(Linear(E_{Pool}))))$$

where σ denotes the sigmoid activation function that transforms the output to a probability value between 0 and 1.

For binding site prediction, the model utilizes local interaction embeddings, $E_{Interactionlocal}$ with a shape of $(L_{RNA}, N_{atom}, 32)$, which capture nucleotide-specific binding information.

$$E_{Interactionlocal} = E_{interaction}[0 : L_{RNA} - 1, 0 : N_{atom} - 1]$$

An important design consideration is that binding sites are only meaningful for positive RNA-compound interactions. Therefore, to ensure consistency between the two prediction tasks, the model incorporates the interaction prediction score as a weighting factor. Specifically, the binding site prediction probabilities are modulated by the interaction probability $P_{Interaction}$, ensuring that binding site predictions are diminished when the interaction probability is low.

The binding site prediction is formulated as:

$$P_{Binding} = \sigma\left(\frac{1}{N_{atom}} \sum_{j=1}^{N_{atom}} Linear(ReLU\right.$$
$$\left.(Linear(E_{Interactionlocal} \times P_{Interaction}))\right)$$

This calculation yields a matrix of shape $(L_{RNA}, 1)$, where each value represents the probability of the corresponding RNA nucleotide being part of the binding site.

Loss functions

For the interaction prediction task, DeepRNA-DTI employs binary cross-entropy loss, a standard choice for binary classification problems.

$$L_{Interaction} = -\big(ylogP_{Interaction} + \big(1 - y\big)log(1 - P_{Interaction})\big)$$

where y represents the ground truth label (1 for positive interactions, 0 for negative interactions) and $P_{Interaction}$ is the predicted interaction probability obtained from the interaction prediction module.

For the binding site prediction task, the model implements focal loss, a modified version of cross-entropy loss designed to address class imbalance in classification problems. The binding site prediction task typically

exhibits significant imbalance, as most nucleotides in an RNA sequence do not participate in binding. The focal loss is formulated as:

$$L_{BindingSite} = -\frac{1}{L_{RNA}} \sum_{i=1}^{L_{RNA}} \alpha(1 - P_i)^{\gamma} y_i log P_i$$

where P_i is the predicted binding probability for the nucleotide at position i, y_i is the ground truth binding label for that position, α and γ are hyperparameters that control the contribution of each sample to the loss, and L_{RNA} is the length of the RNA sequence.

The total loss for model training is the simple sum of the interaction prediction loss and the binding site prediction loss:

$$L_{Total} = L_{BindingSite} + L_{Interaction}$$

This combined loss function enables the model to simultaneously learn both tasks, leveraging the complementary information between them to improve overall performance.

Results and analysis

DeepRNA-DTI outperforms baseline models and existing prediction models

To jointly optimize DTI and binding site prediction, we explored several learning strategies. Among these, the pooled strategy achieved the best performance. Because the binding site dataset was relatively smaller, we applied an upsampling based data imbalance mitigation strategy, which was crucial for stable and effective training. Detailed description of methods and results are in Supplementary 2.2 and 3.1.

We evaluated DeepRNA-DTI in all unseen settings and compared it with both machine learning (ML) and deep learning (DL) models (Fig. 2).

The machine learning models can be categorized into two groups. The first group consisted of an existing comparison model, RSApred, which employs linear regression models trained separately for each RNA subtype. The second group included our in-house ML baseline models that use foundation model embeddings to assess the effectiveness of large scale pretrained representations. Specifically, we use RNA-FM embeddings as RNA features and Mole-BERT embeddings as compound features and concatenating these embeddings becomes input of ML algorithms such as logistic regression, SVM (linear), SVM (rbf), random forest and XGBoost.

The DL models can also be categorized into two categories. The first category consists of baseline models originally designed for protein-compound interactions but adapted for RNA-compound interactions: DeepDTA [49], GraphDTA [50], and GraphATT-DTA [51]. In these adaptations, protein sequences were replaced with RNA sequences, and the loss function was modified to binary cross-entropy. In these models, protein sequences are encoded using a 1D Convolutional Neural Network (CNN), and compound representations vary. DeepDTA encodes SMILES strings of compounds using a 1D CNN. GraphDTA encodes compound graphs using a Graph Convolutional Network (GCN), while GraphATT-DTA also uses GCNs but incorporates an attention mechanism to integrate sequence-atom interactions. The second category includes existing RNA-compound interaction comparison models: DeepRSMA [35] and RNAsmol [41]. DeepRSMA utilizes both sequence and 2D graph representations for RNA and compounds. RNAsmol is a sequence based deep learning model that takes RNA sequence and compound graphs as inputs and applies an attention mechanism for interaction prediction.

All models were trained and evaluated on the same datasets as DeepRNA-DTI to ensure fair comparison. As shown in Fig. 2, DeepRNA-DTI consistently achieves high performance across all unseen settings. In the unseen pair setting, DeepRNA-DTI obtained an average AUC of 0.73, outperforming the second best model (one-tailed *t*-test, $p < 0.05$) (Fig. 2a). In the unseen compound and unseen RNA settings, DeepRNA-DTI also achieved statistically significant improvements (one-tailed *t*-test, $p < 0.05$) with AUCs of 0.629 and 0.58, respectively (Fig. 2b, c). Among machine learning models, using foundation model embeddings (RNA-FM and Mole-BERT) with tree-based models improved performance, while linear models showed only marginal improvement. These results indicate that foundation embeddings are effective when combined with models capable of exploiting their nonlinear property. Among deep learning models, compound graph based architectures exhibited slightly higher and more stable performance than compound sequence based models such as DeepDTA, particularly in the unseen compound setting. This trend suggests that modeling molecular topology provides an advantage in generalizing novel chemical scaffolds. Performance decreased for all models in the unseen RNA and unseen both settings, reflecting the intrinsic data scarcity in current RNA-compound datasets rather than model deficiencies. Expanding and diversifying RNA-compound interaction data will be essential for constructing a more reliable and generalizable model.

For RSApred, we utilized the linear regression equations from the original paper and evaluated their performance on our dataset. Because these equations were derived from a different dataset, to ensure a fair comparision, we additionally evaluate the performance on the RSApred dataset and confirmed that DeepRNA-DTI

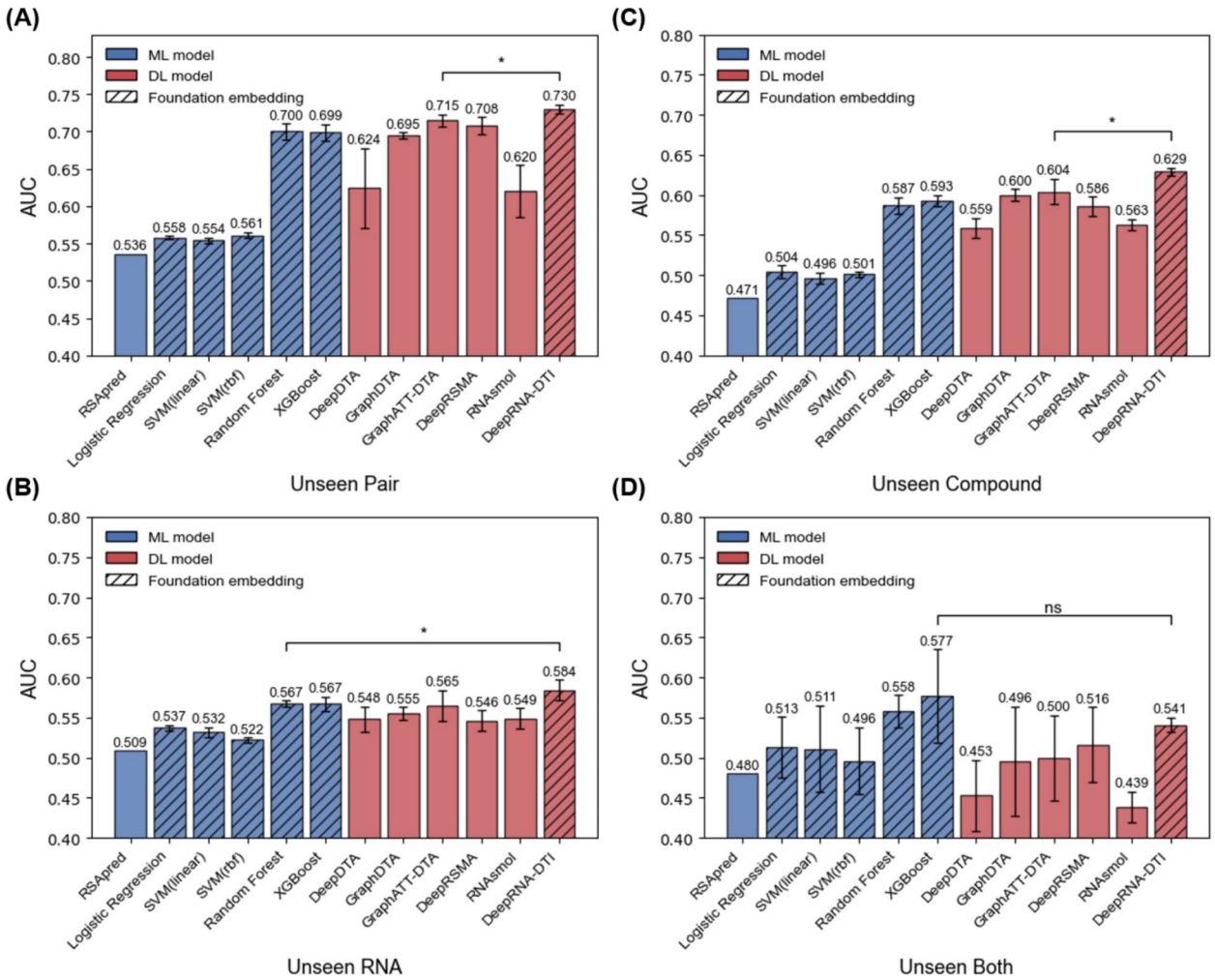

Fig. 2 Performance comparison across different unseen evaluation settings (**A**) Unseen Pair. (**B**) Unseen Compound. (**C**) Unseen RNA. (**D**) Unseen Both

outperforms other models. Detailed data statistics and results are provided in Supplementary 3.2 and 3.3.

Ablation study

To assess the contribution of each component of DeepRNA-DTI, we conducted an ablation study with three settings: (1) training only on the DTI prediction task (No Binding Site), (2) training only on the binding site prediction task (No Interaction), and (3) replacing the attention mechanism with simple concatenation and trained it only on the DTI prediction task (No Attention). For "No Binding Site" and "No Interaction", we retained the multitask architecture but selectively excluded the corresponding training data for each task. Specifically, in the "No Binding Site" setting, the model was trained only on DTI data while maintaining the multitask structure, and was evaluated on both the DTI and binding site test sets to examine cross-task effects. Conversely, in the "No Interaction" setting, the model was trained only on binding site data, preserving the multitask architecture, and was tested on both DTI and binding site test sets. Because both tasks share the same input, the model can predict for both output simultaneously even when only one task is trained.

First, for DTI prediction, if only binding site information was trained without DTI information, the DTI prediction performance dropped significantly (Fig. 3a). However, when binding site information was not included in DTI prediction, we observed that DTI prediction performance also decreased, particularly in challenging cases such as unseen compound, unseen RNA, and unseen both. This result confirms that the binding site prediction task provides knowledge transfer to DTI prediction.

On the other hand, for binding site prediction, we found that performance did not decrease even when DTI information was not jointly trained, and in some cases, it

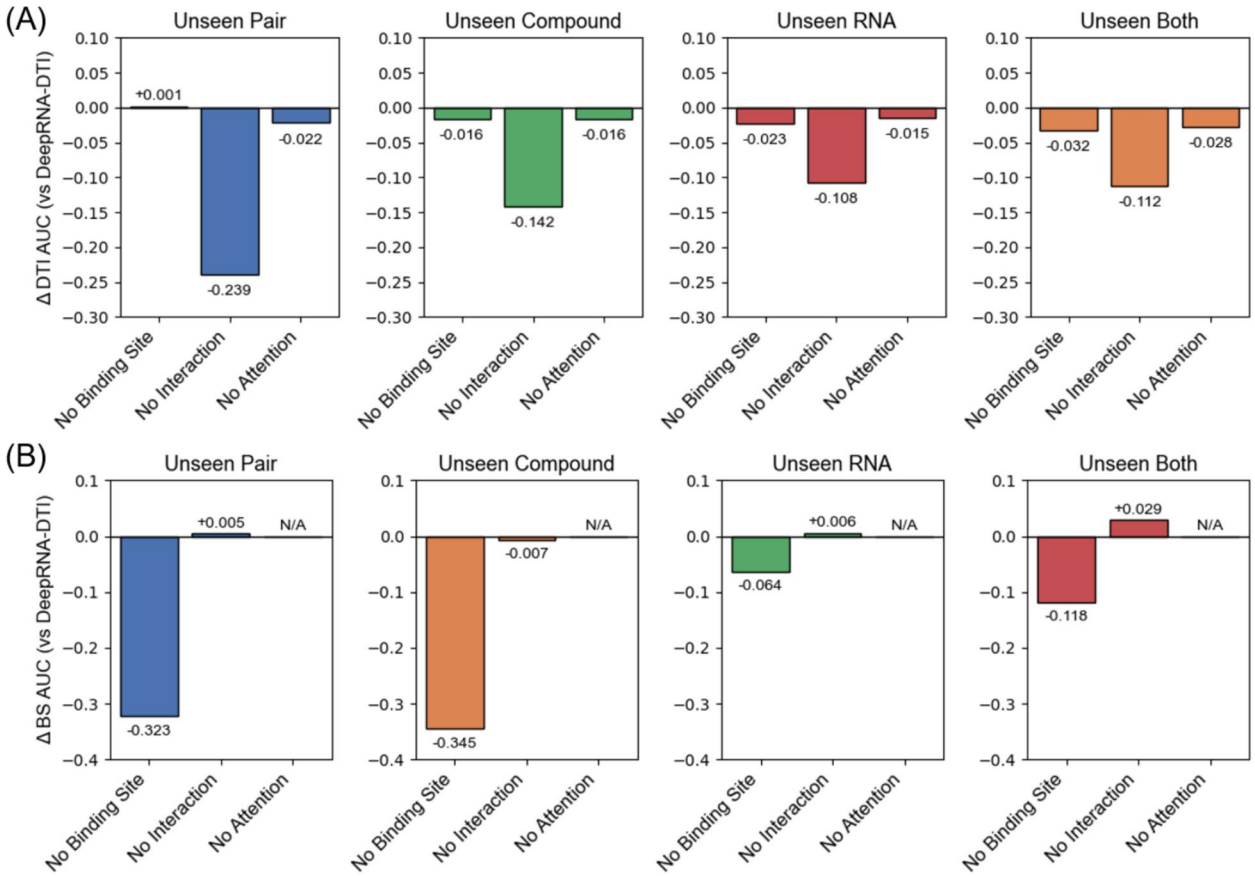

Fig. 3 Ablation Study of DeepRNA-DTI Model Components. **A** Difference in DTI AUC between DeepRNA-DTI and its ablated models. **B** Difference in BS AUC between DeepRNA-DTI and its ablated models

even slightly improved (Fig. 3b). This suggests that binding site information plays a dominant role in binding site prediction, and that DTI information does not provide relevant knowledge transfer to binding site prediction.

In our model, DTI prediction serves as the main task, while binding site prediction is introduced as an auxiliary task to provide nucleotide-level interpretability and enhance DTI prediction through knowledge transfer. As shown in the ablation study results, the joint multitask model achieves improved DTI prediction performance compared to the single-task model, particularly in challenging generalization scenarios, while binding site prediction performance remains comparable or slightly better.

Deep learning based binding site prediction from RNA sequence

Since no previous research has utilized RNA sequences for binding site prediction, we developed two baseline models to evaluate our binding site prediction capabilities. These baseline models were trained and tested on the same dataset as DeepRNA-DTI.

The first baseline, the BS probability model, uses a simple statistical approach. It calculates the binding site ratio (percentage of nucleotides that are binding sites) for each RNA sequence in the test set and predicts binding sites in the test set sequence by randomly selecting nucleotides based on this overall probability with the repetition of 100 times for each RNA.

The second model is a machine learning-based model that uses one hot encoded nucleotide sequence and compound MACCS fingerprints as inputs to predict whether each sequence position is a binding site.

As shown in Fig. 4, DeepRNA-DTI substantially outperforms all baseline models. The purely statistical approach, BS probability, demonstrates poor performance because it relies solely on binding site frequency without considering sequence-specific features or compound information. While the machine learning model achieves marginally better results, it remains limited by its simplistic feature representation and inability to capture complex sequence-compound interaction patterns. In contrast, DeepRNA-DTI achieves superior performance by leveraging deep learning to integrate rich RNA

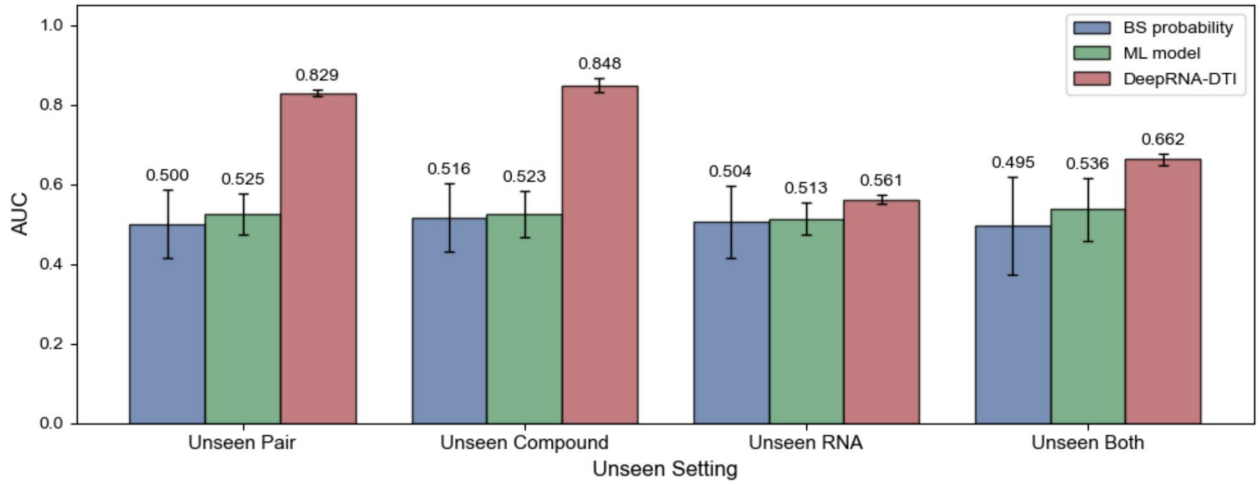

Fig. 4 Performance comparison of Binding site with baseline models

sequence representations from RNA-FM and compound structural information from Mole-BERT. DeepRNA-DTI demonstrates performance improvements in the unseen pair and unseen compound settings, while showing relatively limited performance in unseen RNA and unseen both settings. The observed limitation in unseen RNA and unseen Both cases suggests that the models' generalization remains constrained by the scarcity of available RNA-compound binding site data. These underscore the importance of expanding datasets to enhance generalization across diverse RNA sequences in future research.

To further validate our binding site prediction framework, we compared DeepRNA-DTI with external models, RNAsite and RNABind, and clarified the conceptual distinction between compound-agnostic and compound-specific prediction. Most existing binding site prediction methods are compound-agnostic, taking only RNA sequences as input to identify general binding sites that can interact with any small molecule. In contrast, our model is compound-specific, taking both an RNA sequence and a compound as input. This allows to predict which residues will interact with that particular RNA-compound pair.

We compared its performance with RNAsite [18] and RNABind [26] using a subset of RNA structures that did not overlap with our training data (Table 2). RNAsite achieved an average AUC of 0.743 (std 0.192), RNABind achieved an average AUC of 0.675 (std 0.038), whereas DeepRNA-DTI achieved an average AUC of 0.801 (std 0.189). These results should be interpreted with caution, since RNAsite and RNABind were trained with a 4Å cutoff, whereas we used 10Å. Moreover, RNAsite and RNABind are compound-agnostic and include single atoms such as Mg as ligands, whereas our dataset excludes single atoms. These differences in label definition and data curation affect the comparison, so the result is indicative rather than a rigorous evaluation.

The distinction between compound-specific and compound-agnostic prediction is crucial. A single RNA molecule may contain multiple pockets for different ligands. Compound-agnostic approaches label all of these potential sites as positive binding sites, regardless of the specific ligand being studied. In contrast, the compound-specific approach predicts which residues interact with the given ligand, providing practical value in drug discovery scenarios where the compound of interest is predefined.

In our model, interpretability is defined as the ability to provide a rationale (where it interacts, i.e., the binding site) for the model's decision (if it interacts, i.e., DTI

Table 2 Comparison of existing RNA binding site prediction models

Model	Compound consideration	Binding cutoff	Training dataset	AUC (std)
RNAsite	No (compound agnostic)	4 Å	RNAsite	0.743 (0.192)
RNABind	No (compound agnostic)	4 Å	RNAsite	0.675 (0.038)
DeepRNA-DTI	Yes (compound specific)	10 Å	DeepRNA-DTI (Compound-specific dataset)	0.801 (0.189)

std standard deviation

prediction). The multitask enforces self-consistency between the two predictions. We analyze the Binding Site probability for non-binding pairs in Supplementary 3.4.

Binding site predictions accurately localize to structural binding regions

To evaluate DeepRNA-DTI's predictive power, we examined two challenging test cases that fall outside our training data distribution: one featuring an unseen compound and another with an unseen RNA sequence. For our first test case (Fig. 5a), we analyzed complex 8XZM from the PDB database, which contains a compound highly dissimilar to any in our training set (Tanimoto similarity < 0.3). Despite encountering this chemically novel structure, our model delivered remarkable accuracy when predicting binding sites on the 53-nucleotide RNA, achieving an AUC of 0.985. In our second test case (Fig. 5b), we assessed complex 4Y1M, which features an RNA sequence with limited similarity to our training examples (normalized Needleman-Wunsch score < 0.7). Even with this 107-nucleotide RNA sequence, DeepRNA-DTI performed robustly, yielding an AUC of 0.838.

In addition, our visualizations illustrate the spatial correspondence between predicted and actual binding sites. True binding sites appear in orange on the 3D structure, while predicted binding sites are highlighted in cyan. The accompanying bar plots provide nucleotide-level comparisons, with the x-axis representing the RNA sequence position, and the y-axis showing normalized distance (top) and predicted binding probability (bottom). The computation of normalized distance is detailed in Supplementary 3.5. The substantial overlap between predicted and actual binding site distributions in both cases demonstrates DeepRNA-DTI's ability to make accurate predictions even for novel RNA-compound pairs.

Massive AI screening reveals novel pre-miR-21 targeting scaffolds

To demonstrate DeepRNA-DTI's practical utility in drug discovery, we conducted a large-scale virtual screening targeting pre-miR-21, an oncogenic microRNA that regulates key tumor suppressor genes, including PDCD4 and PTEN. Inhibition of miR-21 represents a promising therapeutic strategy for various cancers, as mature miR-21 silences these tumor suppressor genes. While antisense oligonucleotides have been explored to target the miR-21, significant toxicity concerns remain. Small molecules that inhibit miR-21 biogenesis therefore offer a promising alternative with potentially improved pharmacological properties [1, 35].

We examined our dataset composition and identified that pre-miR-21 sequences were present in both our training and test sets. Our training data included interactions between five pre-miR-21 sequences and 177 compounds, while our test set contained one pre-miR-21 sequence with 36 compounds. When evaluating DeepRNA-DTI's performance specifically on these pre-miR-21 test interactions, the model demonstrated robust discriminative ability with an AUC of 0.773. The model effectively distinguished between binding and non-binding compounds, assigning substantially higher scores to positive interactions (average DTI score: 0.687) compared to negative interactions (average DTI score: 0.353).

To further validate the model's generalizability, we tested four experimentally confirmed pre-miR-21 binding compounds reported by Shortridge et al. [52]. The detailed RNA sequence and compound structures are provided in Supplementary 3.6. All four compounds received DTI scores above 0.8, despite their structural novelty relative to our training compounds. This result demonstrates DeepRNA-DTI's ability to identify true binders beyond its training distribution.

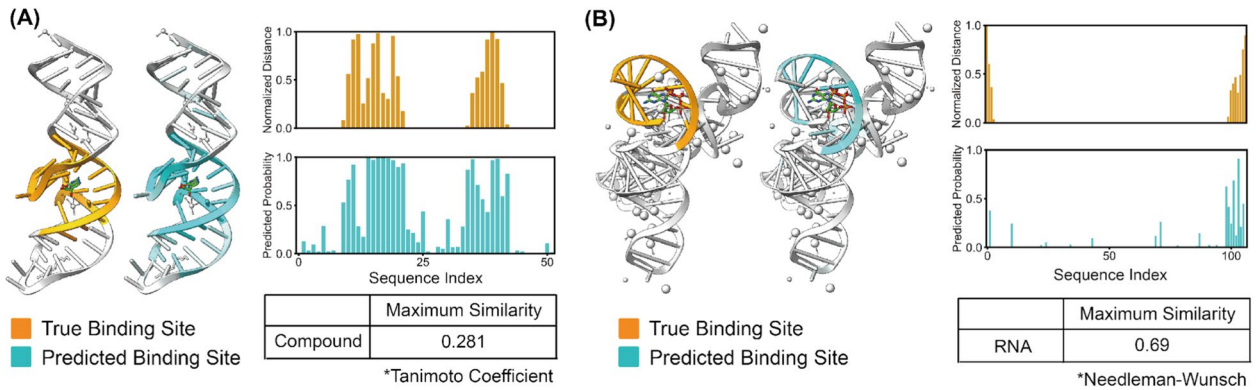

Fig. 5 Visualization of predicted versus actual binding site. **A** Binding Site prediction for low compound similarity (PDB ID: 8XZM). **B** Binding Site prediction for low sequence similarity (PDB ID: 4Y1M)

We then conducted a comprehensive virtual screening of the Enamine REAL diversity library (https://enamine.net/compound-collections/real-compounds) containing 48,232,772 compounds. We selected the top 15,000 compounds with the highest predicted DTI scores for further analysis. This specific number was chosen to match the size of our protein-binding compound reference set, enabling balanced and statistically robust comparisons in subsequent analyses. To characterize the chemical space of these predicted binders, we established five distinct reference groups for comparison: (1) randomly selected Enamine compounds, (2) protein-binding compounds from Lee et al. [53], (3) known RNA-binding compounds, (4) experimentally confirmed pre-miR-21 binding compounds, and (5) our predicted pre-miR-21 binding compounds.

Chemical space visualization using TMAP projection (Fig. 6a) revealed distinct distribution patterns among these compound categories. Each color in the figure corresponds to a compound group, gray (random), orange

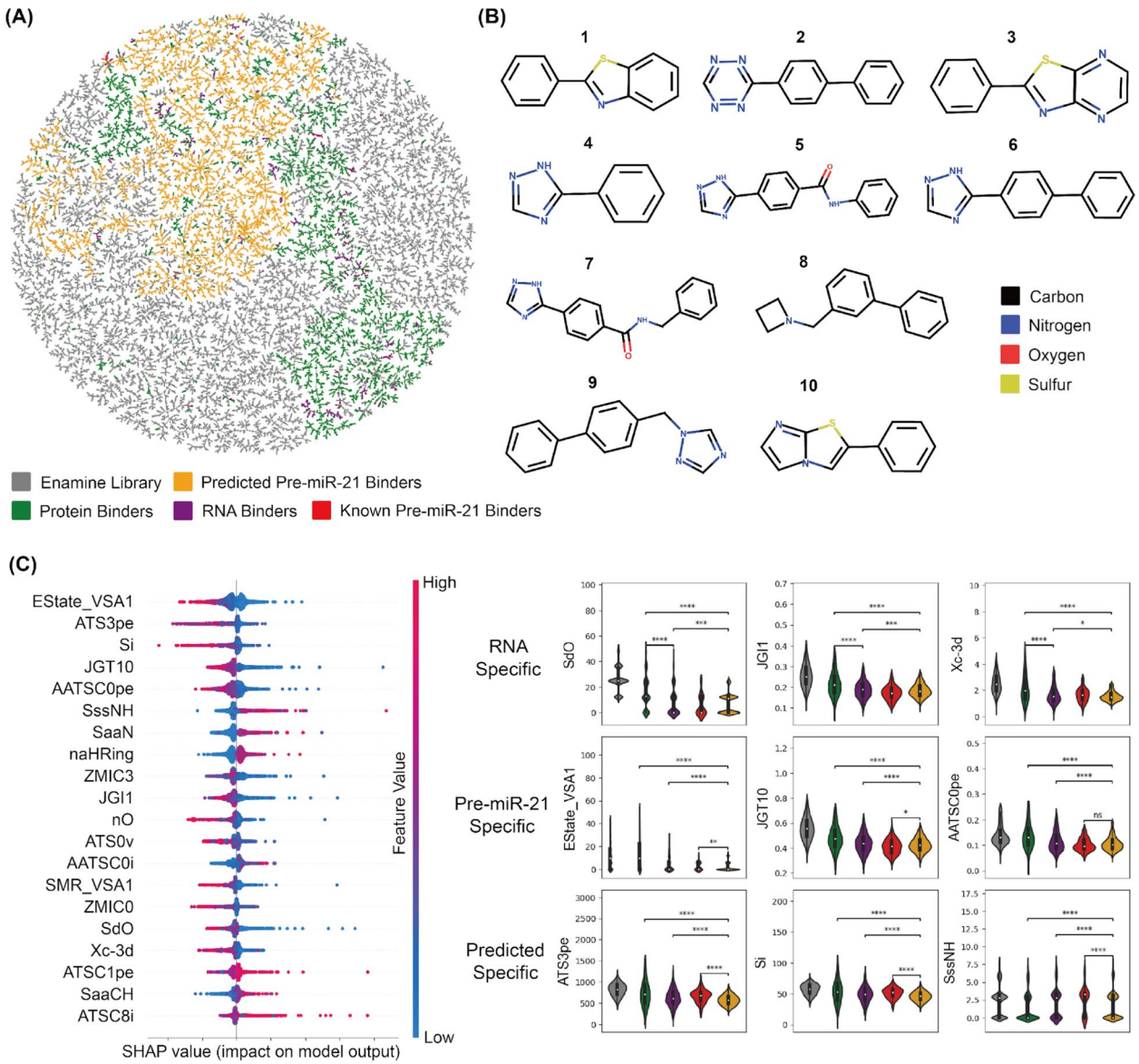

Fig. 6 High Throughput AI Screening for pre-miR-21 binding compounds. **A** Chemical space distribution of compounds. **B** Top 10 most frequently occurring scaffolds among the top 15,000 predicted compounds. **C** Property Analysis of predicted versus known binders. SHAP analysis shows the most informative descriptors distinguishing predicted RNA binders from protein binders (Left), Box plots with corresponding statistical comparison of feature distributions identify RNA-specific, pre-miR-21-specific, and prediction-specific properties (Right)

(predicted Pre-miR-21 binder), green (protein binder), purple (RNA binder) and red (known pre-miR-21 binder). The predicted pre-miR-21 binders occupy specific regions of chemical space that partially overlap with known RNA binders while remaining clearly differentiated from typical protein-binding compounds. To quantitatively assess the chemical space positioning of our predicted compounds, we analyzed the 1-nearest-neighbor of the predicted pre-miR-21 binders. Since the reference sets differ substantially in size (RNA binders: 2529 compounds; protein binders: 14,364 compounds), we randomly subsampled the protein binders to match the RNA binder set size and repeated this analysis ten times to ensure statistical robustness. The results showed that, on average, 61.5% of the predicted pre-miR-21 binders had a RNA binder as their nearest neighbor, whereas 38.5% had a protein binder. This distribution pattern suggests that our model has successfully captured characteristic features associated with RNA-binding molecules rather than general bioactive compounds.

We further examined the chemical scaffolds of the top 15,000 predicted binders using Bemis Murcko scaffold analysis. The top 10 most frequently occurring scaffolds are shown in Fig. 6b. Interestingly, the most frequent scaffold (Scaffold 1) corresponds to a known pre-miR-21 binder, validating that our screening was able to rediscover a known active chemotype. The other nine scaffolds appear to be novel, implying that our model suggests new chemical scaffolds that could be explored as potential pre-miR-21 inhibitors, thereby expanding the pool of candidate chemotypes for RNA-targeted drug design.

To characterize the distinctive chemical properties of our predicted RNA-binding compounds, we conducted a comprehensive molecular descriptor analysis. Initially, we generated a diverse set of molecular descriptors using Mordred [54]. We then trained a Random Forest classifier to distinguish RNA binders from protein binders, and employed SHAP (SHapley Additive exPlanations) analysis [55] to identify the 20 most significant molecular descriptors for RNA binders (Fig. 6c, left). Positive SHAP values indicate that a feature contributes toward the prediction of pre-miR-21 binding, whereas negative values suggest a contribution toward protein binding. Notably, descriptors related to charge, electronic structure, aromaticity, and nitrogen containing substructures emerged as key determinants of pre-miR-21 binding propensity.

For additional precision, we performed Mann–Whitney U statistical tests comparing predicted pre-miR-21 binders against protein binders, organizing the resulting features into three functional categories: RNA-specific, pre-miR-21-specific, and prediction-specific features (Fig. 6c, right). The RNA-specific features we identified include SdO (related to double bonded oxygen), JGI1 (local charge transfer), and Xc-3d (local atomic connectivity and electron distribution), which correspond to established RNA-binding mechanisms involving electrostatic interactions and pi-pi stacking [52, 56, 57]. For pre-miR-21-specific binding, key features include EState_VSA1 (electron density surface area), JGT10 (long range charge effect), and AATSC0pe (localized charge variation), highlighting the distinct electronic and polarity requirements for targeting this specific RNA structure. We also identified prediction-specific features like ATS3pe (electronegativity at local distance), Si (ionization tendency), and SssNH (contribution of NH groups), which could represent additional molecular characteristics potentially enhancing binding effectiveness. Comprehensive statistical analyses and detailed feature comparisons are available in Supplementary 3.6.

Discussion and conclusion

Targeting RNA molecules offers a promising opportunity to expand the therapeutic landscape beyond traditional protein targets, opening new avenues in drug discovery. However, identifying small molecules that bind to RNA have a significant challenge, primarily due to two factors. First, the scarcity of RNA-compound interaction data restricts the development of robust predictive models. Second, existing sequence-based prediction approaches often lack interpretability, hindering mechanistic understanding of RNA-compound interactions.

In this study, we introduce DeepRNA-DTI, a novel deep learning framework designed to predict RNA-compound interactions while simultaneously providing binding site interpretability. Our approach addresses several key challenges in the field.

First, data integration and enhancement. To overcome data scarcity, we curated and integrated multiple datasets from diverse sources, including the PDB and literature, creating a comprehensive collection of RNA-compound interactions. This integration enables more robust model training and broader applicability across RNA subtypes.

Second, sequence-based prediction with structural insights. Unlike structure-dependent approaches, DeepRNA-DTI operates directly on RNA sequences, eliminating the need for 3D structural information. This is particularly advantageous given that 3D structures are unavailable for many RNAs, and even when available, RNA molecules exhibit inherent flexibility and conformational dynamics that are difficult to capture in static models. By leveraging pretrained embeddings from RNA-FM for RNA sequences and Mole-BERT for compounds, our model captures rich sequence context while maintaining structural awareness.

Third, interpretability through multitask learning. DeepRNA-DTI uniquely combines drug-target

interaction prediction with binding site identification in a multitask learning framework. This approach not only improves predictive performance but also provides valuable mechanistic insights into how and where compounds interact with RNA molecules. The joint prediction reflects the biological coupling between nucleotide-level binding and compound activity, enhancing model interpretability without compromising accuracy.

Lastly, balanced training strategy. Our comprehensive evaluation of learning strategies demonstrated that pooled learning with binding site oversampling offers the most effective approach for multi-task prediction. This strategy prevents catastrophic forgetting while enabling beneficial knowledge transfer between tasks, resulting in superior performance for both interaction and binding site predictions.

Performance evaluations demonstrate that DeepRNA-DTI outperforms existing models across multiple datasets. When applied to the RSApred dataset, DeepRNA-DTI achieved an AUC of 0.949, surpassing specialized models like RSApred. This improvement can be attributed to the model's rich pretrained sequence embeddings and attention-based architecture that effectively captures complex interaction patterns. Moreover, despite not being specifically trained for individual RNA subtypes, DeepRNA-DTI shows consistently strong performance across diverse RNA classes, highlighting its robust generalization capabilities.

The practical utility of DeepRNA-DTI was demonstrated through a large-scale virtual screening application targeting pre-miR-21, an oncogenic microRNA implicated in various cancers. The model successfully identified known binders and proposed novel chemical scaffolds from the Enamine REAL diversity library containing over 48 million compounds.

DeepRNA-DTI's sequence-based approach also opens possibilities for exploring therapeutic applications beyond traditional binding inhibition. Small molecules that bind to RNA can modulate RNA function, stability, or interactions in various ways, such as inhibiting oncogenic RNAs or regulating RNA processing events. Our approach may support emerging RNA-targeted therapeutic strategies, including RIBOTACs [58], RNATACs, or RATACs [59], which employ targeted degradation mechanisms similar to proteolysis-targeting chimeras (PROTACs) but directed at RNA molecules.

Although DeepRNA-DTI offers several advantages, it still has some limitations. First, the model outputs point predictions without quantifying uncertainty, which is crucial in virtual screening. Prioritizing compounds with both high predicted scores and high confidence can improve experimental efficiency. This limitation could be addressed using uncertainty aware approaches such as ensemble or Monte Carlo dropout [60, 61]. Second, for binding site prediction, only the ligand closest to the RNA was used when multiple identical ligands were present in a single structure. This simplification may exclude relevant binding information. Incorporating all relevant ligand binding sites could provide more accurate supervision and enhance prediction quality. We also compared DeepRNA-DTI with RNAsite and RNABind. While DeepRNA-DTI showed higher performance, the comparison should be interpreted cautiously due to differences in cutoff definitions and ligand treatment. Additionally, unlike existing compound-agnostic predictors that identify general binding sites, DeepRNA-DTI performs compound-specific prediction for individual RNA-ligand pairs, directly predicting which residues are involved in binding the given ligand. While this design aligns with drug discovery scenarios where the compound of interest if predefined, extending the framework to support both compound-specific and compound-agnostic binding site prediction would be valuable for broader applications.

In conclusion, DeepRNA-DTI demonstrates that a sequence-based deep learning approach can achieve accurate and interpretable predictions of RNA-compound interactions. By combining rich sequence representations with multitask learning for simultaneous interaction and binding site prediction, our model provides valuable insights into RNA-compound binding mechanisms while maintaining high predictive performance. This work represents a significant step toward leveraging computational methods for RNA-targeted small molecule discovery and could accelerate the development of novel therapeutics against previously "undruggable" targets.

Abbreviations

DTI	Drug Target Interaction
pKd	Negative logarithm of the dissociation constant (Kd)
SMILES	Simplified molecular-input line-entry system
AUC	Area under the curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01132-y.

Additional file 1.

Author contributions

HB and HN conceptualized the study. HB implemented the model. HB performed experiments. HB prepared the initial draft. HN revised the manuscript and supervised the study.

Funding

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2025-02304253), and supported by a grant of the National Research Foundation of Korea (NRF) grant funded by the Korea

government (MSIT) (RS-2024-00334990). This work was supported by Artificial Intelligence Industrial Convergence Cluster Development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

Data availability

The data and source code for DeepRNA-DTI are available at our GitHub repository ([https://github.com/GIST-CSBL/DeepRNA-DTI/] (https:/github.com/GIST-CSBL/DeepRNA-DTI)) under the PolyForm Noncommercial License 1.0.0. All trained model weights and generated data using our source code are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]AI Graduate School, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea. [2]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea. [3]Center for AI-Applied High Efficiency Drug Discovery (AHEDD), Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea.

Received: 5 August 2025 Accepted: 26 November 2025
Published online: 02 December 2025

References

1. Kaur J, Sharma A, Mundlia P, Sood V, Pandey A, Singh G, Barnwal RP (2024) RNA–small-molecule interaction: challenging the "undruggable" tag. J Med Chem 67:4259–4297. https://doi.org/10.1021/acs.jmedchem.3c01354
2. Childs-Disney JL, Yang X, Gibaut QMR, Tong Y, Batey RT, Disney MD (2022) Targeting RNA structures with small molecules. Nat Rev Drug Discov 21:736–762. https://doi.org/10.1038/s41573-022-00521-4
3. Warner KD, Hajdin CE, Weeks KM (2018) Principles for targeting RNA with drug-like small molecules. Nat Rev Drug Discov 17:547–558. https://doi.org/10.1038/nrd.2018.93
4. Winkle M, El-Daly SM, Fabbri M, Calin GA (2021) Noncoding RNA therapeutics—challenges and potential solutions. Nat Rev Drug Discov 20:629–651. https://doi.org/10.1038/s41573-021-00219-z
5. Kovachka S, Panosetti M, Grimaldi B, Azoulay S, Di Giorgio A, Duca M (2024) Small molecule approaches to targeting RNA. Nat Rev Chem 8:120–135. https://doi.org/10.1038/s41570-023-00569-9
6. Chen S, Mao Q, Cheng H, Tai W (2024) RNA-binding small molecules in drug discovery and delivery: an overview from fundamentals. J Med Chem 67:16002–16017. https://doi.org/10.1021/acs.jmedchem.4c01330
7. Li M, Zhu J, Lv Z, Qin H, Wang X, Shi H (2024) Recent advances in RNA-targeted cancer therapy. ChemBioChem 25:e202300633. https://doi.org/10.1002/cbic.202300633
8. Nemeth K, Bayraktar R, Ferracin M, Calin GA (2024) Non-coding RNAs in disease: from mechanisms to therapeutics. Nat Rev Genet 25:211–232. https://doi.org/10.1038/s41576-023-00662-1
9. Chen L-L, Kim VN (2024) Small and long non-coding RNAs: past, present, and future. Cell 187:6451–6485. https://doi.org/10.1016/j.cell.2024.10.024
10. Chen B, Dragomir MP, Yang C, Li Q, Horst D, Calin GA (2022) Targeting non-coding RNAs to overcome cancer therapy resistance. Signal Transduct Target Ther 7:121. https://doi.org/10.1038/s41392-022-00975-3
11. Yu A-M, Choi YH, Tu M-J (2020) RNA drugs and RNA targets for small molecules: principles, progress, and challenges. Pharmacol Rev 72:862–898. https://doi.org/10.1124/pr.120.019554
12. Fullenkamp CR, Schneekloth JS (2023) RNA as an off-target for FDA-approved drugs. Nat Chem 15:1329–1331. https://doi.org/10.1038/s41557-023-01330-x
13. Costales MG, Childs-Disney JL, Haniff HS, Disney MD (2020) How we think about targeting RNA with small molecules. J Med Chem 63:8880–8900. https://doi.org/10.1021/acs.jmedchem.9b01927
14. Falese P, Donlic J, Hargrove AE (2021) Targeting RNA with small molecules: from fundamental principles towards the clinic. Chem Soc Rev 50:2224–2243. https://doi.org/10.1039/D0CS01261K
15. Lightfoot HL, Smith GF (2025) Targeting RNA with small molecules—a safety perspective. Br J Pharmacol. https://doi.org/10.1111/bph.16027
16. Morishita EC, Nakamura S (2024) Recent applications of artificial intelligence in RNA-targeted small molecule drug discovery. Expert Opin Drug Discov 19:415–431. https://doi.org/10.1080/17460441.2024.2313455
17. Zhou Y, Chen S-J (2024) Advances in machine-learning approaches to RNA-targeted drug design. Artif Intell Chem 2:100053. https://doi.org/10.1016/j.aichem.2024.100053
18. Su H, Peng Z, Yang J (2021) Recognition of small molecule–RNA binding sites using RNA sequence and structure. Bioinformatics 37:36–42. https://doi.org/10.1093/bioinformatics/btaa1092
19. Jiang Z, Xiao S-R, Liu R (2022) Dissecting and predicting different types of binding sites in nucleic acids based on structural information. Brief Bioinform 23:bbab411. https://doi.org/10.1093/bib/bbab411
20. Wang K, Zhou R, Wu Y, Li M (2023) RLBind: a deep learning method to predict RNA–ligand binding sites. Brief Bioinform 24:bbac486. https://doi.org/10.1093/bib/bbac486
21. Chhabra S, Xie J, Frank AT (2020) RNAPosers: machine learning classifiers for ribonucleic acid-ligand poses. J Phys Chem B 124:4436–4445. https://doi.org/10.1021/acs.jpcb.0c02322
22. Kozlovskii I, Popov P (2021) Structure-based deep learning for binding site detection in nucleic acid macromolecules. NAR Genomics Bioinform 3:lqab111. https://doi.org/10.1093/nargab/lqab111
23. Stefaniak F, Bujnicki JM (2021) AnnapuRNA: a scoring function for predicting RNA-small molecule binding poses. PLoS Comput Biol 17:e1008309. https://doi.org/10.1371/journal.pcbi.1008309
24. Xiao H, Yang X, Zhang Y, Zhang Z, Zhang Ge, Zhang B-T (2023) RNA-targeted small-molecule drug discoveries: a machine-learning perspective. RNA Biol 20:384–397. https://doi.org/10.1080/15476286.2023.2223498
25. Xi W, Wang R, Wang L, Ye X, Liu M, Sakurai T (2024) An interpretable deep learning model predicts RNA–small molecule binding sites. Future Gener Comput Syst 159:557–566. https://doi.org/10.1016/j.future.2024.05.029
26. Zhu W, Ding X, Shen H-B, Pan X (2025) Identifying RNA-small molecule binding sites using geometric deep learning with language models. J Mol Biol 437:169010. https://doi.org/10.1016/j.jmb.2025.169010
27. Disney MD, Winkelsas AM, Velagapudi SP, Southern M, Fallahi M, Childs-Disney JL (2016) Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured RNAs. ACS Chem Biol 11:1720–1728. https://doi.org/10.1021/acschembio.6b00001
28. Donlic A, Swanson EG, Chiu L-Y, Wicks SL, Juru AU, Cai Z, Kassam K, Laudeman C, Sanaba BG, Sugarman A, Han E, Tolbert BS, Hargrove AE (2022) R-BIND 2.0: an updated database of bioactive RNA-targeting small molecules and associated RNA secondary structures. ACS Chem Biol 17:1556–1566. https://doi.org/10.1021/acschembio.2c00224
29. Sun S, Yang J, Zhang Z (2022) RNALigands: a database and web server for RNA–ligand interactions. RNA 28:115–122. https://doi.org/10.1261/rna.078889.121
30. Oliver C, Mallet V, Gendron RS, Reinharz V, Hamilton WL, Moitessier N, Waldispühl J (2020) Augmented base pairing networks encode RNA-small molecule binding preferences. Nucleic Acids Res 48:7690–7699. https://doi.org/10.1093/nar/gkaa583
31. Cai Z, Zafferani M, Akande OM, Hargrove AE (2022) Quantitative structure-activity relationship (QSAR) study predicts small-molecule binding to RNA structure. J Med Chem 65:7262–7277. https://doi.org/10.1021/acs.jmedchem.2c00254
32. Yazdani K, Jordan D, Yang M, Fullenkamp CR, Calabrese DR, Boer R, Hilimire T, Allen TEH, Khan RT, Schneekloth JJS (2023) Machine learning informs RNA-binding chemical space. Angew Chem 135:e202211358. https://doi.org/10.1002/ange.202211358
33. Krishnan SR, Roy A, Gromiha MM (2024) Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. Brief Bioinform 25:bbae002. https://doi.org/10.1093/bib/bbae002
34. Sun S, Gao L (2024) Contrastive pre-training and 3D convolution neural network for RNA and small molecule binding affinity prediction. Bioinformatics 40:btae155. https://doi.org/10.1093/bioinformatics/btae155
35. Huang Z, Wang Y, Chen S, Tan YS, Deng L, Wu M (2024) DeepRSMA: a cross-fusion-based deep learning method for RNA–small molecule

binding affinity prediction. Bioinformatics 40:btae678. https://doi.org/10.1093/bioinformatics/btae678
36. Zhuo C, Gao J, Li A, Liu X, Zhao Y (2024) A machine learning method for RNA–small molecule binding preference prediction. J Chem Inf Model 64:7386–7397. https://doi.org/10.1021/acs.jcim.4c01324
37. Carvajal-Patiño JG, Mallet V, Becerra D, Niño Vasquez LF, Oliver C, Waldispühl J (2025) RNAmigos2: accelerated structure-based RNA virtual screening with deep graph learning. Nat Commun 16:2799. https://doi.org/10.1038/s41467-025-57852-0
38. Xia Y, Li J, Chu Y-T, Rao J, Chen J, Hua C, Yu D-J, Chen X-C, Zheng S (2025) GerNA-Bind: Geometric-enhanced RNA-ligand Binding Specificity Prediction with Deep Learning. 2025.02.15.638393
39. Tahmid MT, Abir AR, Bayzid MdS (2024) BioLLMNet: enhancing RNA-interaction prediction with a specialized cross-LLM transformation network
40. Wu Z, Deng Z, Wu Q, Zuo Y, Pan X, Shen H, Fang X, Ge Y, Hu S (2025) Explainable RNA-small molecule binding affinity prediction based on multiview enhancement learning. J Chem Inf Model 65:6397–6416. https://doi.org/10.1021/acs.jcim.5c01064
41. Ma H, Gao L, Jin Y, Ma J, Bai Y, Liu X, Bao P, Liu K, Xu ZZ, Lu ZJ (2025) RNA–ligand interaction scoring via data perturbation and augmentation modeling. Nat Comput Sci 5:648–660. https://doi.org/10.1038/s43588-025-00820-x
42. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM (2019) The roles of structural dynamics in the cellular functions of RNAs. Nat Rev Mol Cell Biol 20:474–489. https://doi.org/10.1038/s41580-019-0136-0
43. Umuhire Juru A, Patwardhan NN, Hargrove AE (2019) Understanding the contributions of conformational changes, thermodynamics, and kinetics of RNA–small molecule interactions. ACS Chem Biol 14:824–838. https://doi.org/10.1021/acschembio.8b00945
44. Zhuo C, Zeng C, Yang R, Liu H, Zhao Y (2023) RPflex: a coarse-grained network model for RNA pocket flexibility study. Int J Mol Sci 24:5497. https://doi.org/10.3390/ijms24065497
45. Szulc NA, Mackiewicz Z, Bujnicki JM, Stefaniak F (2023) Structural interaction fingerprints and machine learning for predicting and explaining binding of small molecule ligands to RNA. Brief Bioinform 24:bbad187. https://doi.org/10.1093/bib/bbad187
46. Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, Shen T, King I, Li Y (2022) Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions
47. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res 49:D212–D220. https://doi.org/10.1093/nar/gkaa921
48. Xia J, Zhao C, Hu B, Gao Z, Tan C, Liu Y, Li S, Li SZ (2023) Mole-BERT: rethinking pre-training graph neural networks for molecules
49. Öztürk H, Özgür A, Ozkirimli E (2018) DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 34:i821–i829. https://doi.org/10.1093/bioinformatics/bty593
50. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S (2021) GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 37:1140–1147. https://doi.org/10.1093/bioinformatics/btaa921
51. Bae H, Nam H (2023) GraphATT-DTA: attention-based novel representation of interaction to predict drug-target binding affinity. Biomedicines 11:67. https://doi.org/10.3390/biomedicines11010067
52. Shortridge MD, Chaubey B, Zhang HJ, Pavelitz T, Vidadala V, Tang C, Olsen GL, Calin GA, Varani G (2023) Drug-like small molecules that inhibit expression of the oncogenic MicroRNA-21. ACS Chem Biol 18:237–250. https://doi.org/10.1021/acschembio.2c00502
53. Lee I, Nam H (2022) Sequence-based prediction of protein binding regions and drug–target interactions. J Cheminform 14:5. https://doi.org/10.1186/s13321-022-00584-w
54. Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminform 10:4. https://doi.org/10.1186/s13321-018-0258-y
55. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.
56. Umuhire Juru A, Hargrove AE (2021) Frameworks for targeting RNA with small molecules. J Biol Chem 296:100191. https://doi.org/10.1074/jbc.REV120.015203
57. Suresh BM, Li W, Zhang P, Wang KW, Yildirim I, Parker CG, Disney MD (2020) A general fragment-based approach to identify and optimize bioactive ligands targeting RNA. Proc Natl Acad Sci USA 117:33197–33203. https://doi.org/10.1073/pnas.2012217117
58. Dey SK, Jaffrey SR (2019) RIBOTACs: small molecules target RNA for degradation. Cell Chem Biol 26:1047–1049. https://doi.org/10.1016/j.chembiol.2019.07.015
59. Song Y, Cui J, Zhu J, Kim B, Kuo M-L, Potts PR (2024) RNATACs: multispecific small molecules targeting RNA by induced proximity. Cell Chem Biol 31:1101–1117. https://doi.org/10.1016/j.chembiol.2024.05.006
60. Kim Qh, Ko J-H, Kim S, Park N, Jhe W (2021) Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. Bioinformatics 37:3428–3435. https://doi.org/10.1093/bioinformatics/btab346
61. Luo Y, Liu Y, Peng J (2023) Calibrated geometric deep learning improves kinase–drug binding predictions. Nat Mach Intell 5:1390–1401. https://doi.org/10.1038/s42256-023-00751-0

Publisher's Note