Nucleic Acids Research, 2025, **53**, gkaf239
https://doi.org/10.1093/nar/gkaf239
Methods

DRLiPS: a novel method for prediction of druggable RNA-small molecule binding pockets using machine learning

Sowmya Ramaswamy Krishnan [1,2], **Arijit Roy** [2,*], **Limsoon Wong** [3], **M. Michael Gromiha** [1,3,*]

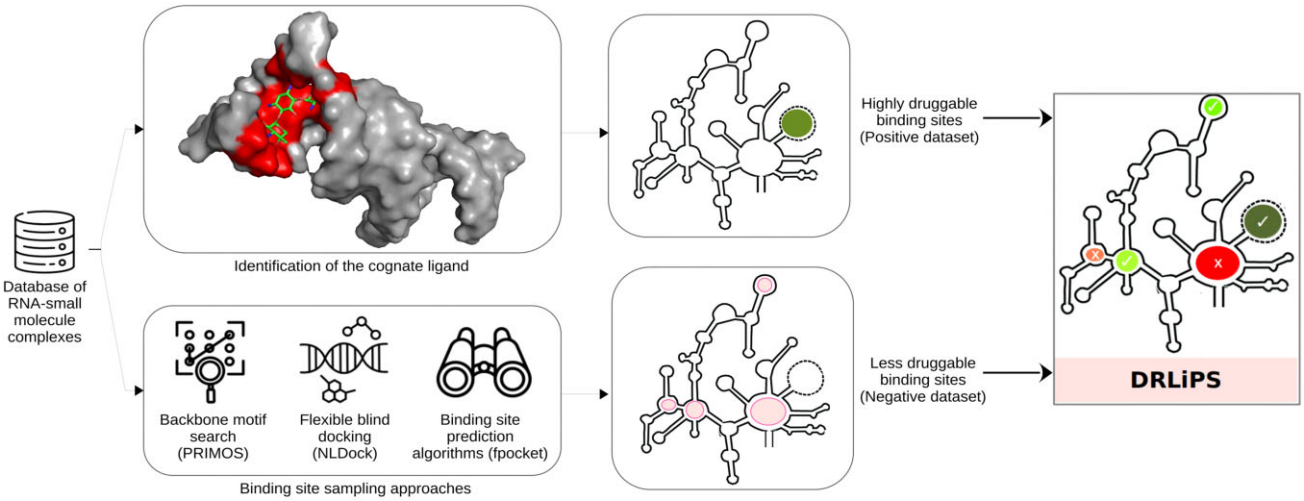
[1]Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India
[2]TCS Research (Life Sciences division), Tata Consultancy Services, Hyderabad 500081, India
[3]Department of Computer Science, National University of Singapore, 117417, Singapore

[*]To whom correspondence should be addressed. Email: gromiha@iitm.ac.in
Correspondence may also be addressed to Arijit Roy. Email: roy.arijit3@tcs.com

Abstract

Ribonucleic Acid (RNA) is the central conduit for information transfer in the cell. Identifying potential RNA targets in disease conditions is a challenging task, given the vast repertoire of functional non-coding RNAs in a human cell. A potential druggable target must satisfy several criteria, including disease association, cellular accessibility, binding pockets for drug-like molecules, and minimal cross-reactivity. While several methods exist for prediction of druggable proteins, they cannot be repurposed for RNAs due to fundamental differences in their binding modality. Taking all these constraints into account, a new structure-based model, **D**ruggable **R**NA-**Li**gand binding **P**ocket **S**elector (**DRLiPS**), is developed here to predict binding site-level druggability of any given RNA target. A novel strategy for sampling negative binding sites in RNA structures using three parallel approaches is demonstrated here to improve model specificity: backbone motif search, exhaustive pocket prediction, and blind docking. An external blind test dataset has also been curated to showcase the model's generalizability to both experimental and modelled apo state RNA structures. DRLiPS has achieved an F1-score of 0.70, precision of 0.61, specificity of 0.89, and recall of 0.73 on this external test dataset, outperforming two existing methods, DrugPred_RNA and RNACavityMiner. Further analysis indicates that the features selected for model-building generalize well to both apo and holo states with a backbone RMSD tolerance of 3 Å. It can also predict the effect of binding site single point mutations on druggability, which can aid in optimizing synthetic RNA aptamers for small molecule recognition. The DRLiPS model is freely accessible at https://web.iitm.ac.in/bioinfo2/DRLiPS/.

Graphical abstract



Introduction

Ribonucleic Acid (RNA) is the central conduit for information transfer in the cell [1]. RNAs are known to sustain disease conditions through enhanced transcription and translation of disease-associated genes and proteins [2, 3]. They can also suppress translation of anti-apoptotic and DNA

Received: October 2, 2024. **Revised:** February 16, 2025. **Editorial Decision:** March 11, 2025. **Accepted:** March 14, 2025

repair proteins in cancers, hastening disease progression [4]. Consequently, RNA subtypes such as miRNA, piwiRNA, and snoRNA have been used as biomarkers to indicate disease progression in the cell [5, 6]. In modern medicine, synthetic RNAs are being utilized as vaccines and therapeutics to target the undruggable proteins at the transcript level [7, 8]. Notably, the identification of siRNA and the ensuing RNA interference (RNAi) therapy has made translational repression a successful therapeutic modality for cancers [9, 10]. However, despite their roles in multiple disease conditions, only coding RNAs have been probed so far as drug targets themselves, for small molecule discovery [1]. Ribosome-targeted antibiotics are the predominant class of RNA-targeted drugs with FDA approval, followed by the recent approval of Risdiplam [11], targeting a non-coding RNA to treat spinal muscular atrophy (SMA). With more than 100 000 human long non-coding RNAs (lncRNAs) in registry [12] and the prevalence of a dynamic tertiary structure for most of them [13], non-coding RNAs potentially accurately capture the regulatory state of a cell at any given time point [12]. Considering RNAs as drug targets for small molecule discovery can significantly expand the druggable genome space and enable the emergence of better therapeutic options to treat undruggable diseases. However, identifying potential RNA targets in disease conditions is a challenging task, given the vast repertoire of functional non-coding RNAs in a human cell.

Target identification and validation are crucial milestones marking the initiation of the drug design process [14–16]. A potential drug target must satisfy several criteria, including the presence of validated disease associations, ease of cellular accessibility, availability of binding pockets for drug-like molecules, and the ability to elicit the desirable biological response with minimal cross-reactivity [17, 18]. These criteria have been summarized into a single term in traditional literature as 'Druggability' [19, 20]. The druggability of a target biomolecule is primarily ascertained through the similarity of the target to previously known drug targets in the pharma industry [17]. However, this practice has resulted in an exploratory bias in the space of drug targets, with popular protein families such as G-protein coupled receptors, ion channels, kinases, and nuclear hormone receptors being the most explored [21, 22]. To mitigate this bias, studies at a whole cell level, such as functional genomics assays, are performed to establish the genotype–phenotype relationship for novel targets and understand the effect of genetic variants on the disease-associated pathways [23, 24]. In contrast, phenotypic screening with diverse small molecule microarrays helps in the identification of both novel drug targets and starter scaffolds with high binding affinity [25, 26]. While current experimental methods look at the phenotypic relevance of a drug target to a disease, it is also essential to study the global structural features and binding site-level characteristics of the target molecule, to prioritize and validate their interaction with drug-like small molecules. This is a key aspect when RNAs are considered as drug targets due to their highly dynamic nature and their potential to form multiple binding pockets for small molecules [27, 28]. With the available three-dimensional structures of target-small molecule complexes in Protein Data Bank (PDB), multiple computational methods have been developed to quantitatively measure the druggability of a target protein or nucleic acid [16, 29–41].

Existing computational methods for predicting druggability of target proteins can be sub-divided into two categories: whole target-level scoring and binding site-level scoring methods. There are relatively few whole target-level methods in comparison with binding site-level methods. The DrugnomeAI framework [16] and the PINNED method [41] are two recent whole-target level methods, which incorporate disease association, functional significance, interaction with other biomolecules in the cell, and gene expression-based features, to prioritize targets based on their druggability. While both methods incorporate multiple modalities of information to capture a target protein's significance in the cellular and disease context, they cannot precisely pinpoint the binding site for a drug-like small molecule in the target protein. On the other hand, binding site-level scoring methods can take multiple potential binding sites within the same target protein or across different proteins, and discriminate the highly druggable sites of interest. Methods inspired by the biophysical characteristics of ligand binding, such as maximal affinity prediction (MAP_{POD} score) [30] and nuclear magnetic resonance (NMR) hit rate prediction [29, 31], correlate the energetic characteristics of binding sites to their druggability. Methods such as SiteScore [32], DoGSiteScorer [36] and PockDrug [38], are based on the shape and physicochemical characteristics of binding sites. Drug-like density (DLID) [33] and DrugFEATURE [37] are statistical methods which utilize the frequency of drug-like pockets surrounding a target pocket of interest and the similarity of the binding pocket microenvironment to known druggable pockets, respectively, to derive the scoring function. Rule-based methods utilizing a combination of physicochemical properties, such as pocket volume, pocket depth, hydrophobicity, enclosure, and percentage of charged residues, has also been proposed to identify druggable binding pockets [35]. All the above methods utilize standard datasets such as the non-redundant druggable and less druggable (NRDLD) binding sites [34] to train and/or validate their performance.

Similarly, to predict the druggability of RNA targets, only two methods have been developed so far: DrugPred_RNA [39] and RNACavityMiner [40]. Both DrugPred_RNA and RNACavityMiner are binding site-level methods. However, due to the paucity of non-redundant RNA-small molecule complex structures, the generalizability of existing methods to novel RNA structures has not been extensively validated. Specifically, it is necessary for an RNA-specific druggability prediction method to generalize equally well to apo structures, given the magnitude of conformational change expected upon ligand binding to the RNA, as in the case of riboswitches [42, 43].

Considering all these constraints, a new structure-based model, **D**ruggable **R**NA-**Li**gand binding **P**ocket **S**elector (**DRLiPS**), is developed in this paper to predict binding site-level druggability of any given RNA target. The model is trained on a non-redundant dataset of RNA-small molecule complexes from PDB, augmented with a negative dataset obtained through a consensus from three parallel binding site sampling strategies. Multiple machine learning methods were tested and tuned, among which a support vector machine (SVM) method with sigmoid kernel was chosen as the final model. An external blind test dataset was also curated to showcase the model's generalizability to both empirical and modelled apo state RNA structures.

The results indicate that if the root mean square deviation (RMSD) between apo and holo state backbones is within 3 Å, the method can reliably prioritize binding sites for both states

of the RNA. Upon analyzing the effect of single point mutations on model predictions, it was found that the DRLiPS druggability score captures the increase in promiscuity of the binding pockets post mutation in several cases. However, the model could not capture the effect of distal site mutations on the druggability of the cognate binding pocket, as in the case of SAM-I riboswitch [44]. In summary, the analysis indicates that the features selected for building the model generalize well to both empirical and modelled RNA structures in apo and holo states. The model also predicts the effect of binding site single point mutations on druggability, which could aid in optimizing small molecule binding sites in synthetic RNA aptamers [45]. The DRLiPS model can be accessed for prediction at https://web.iitm.ac.in/bioinfo2/DRLiPS/.

Materials and methods

Positive binding site dataset curation

The experimentally determined structures of 1073 RNA-small molecule complexes were curated from PDB (as of 1 June 2024) [46]. The structures were processed to remove cases where the small molecules are stabilizers of RNA structure, such as spermine [47], or when no other small molecules except ions are present, such as the fluorine and manganese riboswitches. Further, structures containing synthetic RNA aptamers were also pruned to account for only naturally occurring RNA targets with established role in disease-associated pathways. After pruning, a dataset of 861 RNA-small molecule complexes were obtained, which were further sub-divided into ribosomal and non-ribosomal RNA complexes. While multiple antibiotic binding sites have been observed within the ribosome, the aminoacyl site (A-site) in the small subunit is the primary druggable site for antibiotic recognition and functional modulation [48, 49]. Hence, out of the ribosomal RNA complexes, 51 complexes containing only the A-site of prokaryotic and eukaryotic ribosomes were included with the non-ribosomal RNA complexes, to obtain a final set of 399 structures as the positive binding site dataset. It is also notable that, all the unique small molecules in complex with the positive dataset are present either in the DrugBank database [50] or the ChEMBL database [51] of drug-like small molecules.

From the dataset of 399 structures, the experimentally observed binding site residues were extracted using a distance cut-off of 6.5 Å between all nucleotide-small molecule heavy atom pairs, following the sc-PDB [52] convention. Any water molecules and ions present within the distance cut-off were also extracted for every binding site. Structures containing multiple copies of the small molecule, such as the NAD + Class II riboswitch [53], contributed more than one binding site to the dataset. To account for the dynamics of RNA structure in solution, unique druggable binding sites were extracted for every structure model deposited from NMR experiments. This accounts for the differences in binding site residue conformations observed between NMR models upon binding the small molecule in solution (Supplementary Section S1 and Supplementary Fig. S1). Detailed statistics about the positive dataset is provided in the 'Results and discussion' section.

Resolving redundancy in the positive dataset

From the curated positive dataset, multiple RNAs were observed with the same cognate binding site capable of interacting with multiple ligands with different affinities, such as the interaction between guanine riboswitch and purine analogs. To resolve this redundancy in the binding sites, initially the RNA targets were clustered based on their Rfam family assignment [54]. The Rfam family for each RNA target was identified through the Infernal *cmscan* program [55], with the PDB-derived RNA sequence as the input. The remaining PDB structures were subject to manual classification based on the structure title and structure-associated literature. In contrast, multiple different binding sites for the same cognate ligand were also observed for the same RNA target, such as the tetrahydrofolate (THF) riboswitch. In these cases, each binding site of the RNA was considered unique to account for the differences in their binding microenvironments [37]. For every Rfam family in the positive dataset, the PDB structures with druggable binding sites were collected and the average value of each feature was computed for the family. In case of NMR structures, the features were first computed for each model and the average feature vector was used for the training dataset. The distributions of each averaged feature obtained for the positive dataset were extensively compared with that of the negative dataset to understand if they are significantly different (based on Kolmogorov–Smirnov test with $P < 0.05$).

Negative binding site dataset curation

An extensive literature survey revealed the unavailability of a standard negative dataset for ligandability or druggability prediction at binding site level for RNA targets. To address this gap, three strategies were developed, drawing on existing approaches for curating negative datasets for protein druggability prediction [19], as explained below.

(a) **Backbone motif search for non-selective binding sites**: Based on studies probing the prominent intermolecular interactions observed in experimental RNA-small molecule complexes, backbone–ligand interactions have been shown to account for less than 10% of ligand selectivity to the RNA target [56]. To leverage this aspect, a comprehensive backbone motif search was performed using the PRIMOS pseudo-torsion matching program [57], with a database of input RNA backbones constructed using the positive binding site dataset. Any backbone matches obtained with less than 2.5 Å RMSD, containing at least four residues, and involving pairs of unrelated RNA families, were considered to be potential negative binding sites for druggability prediction. This can also be justified by the observation that, successful target engagement could not be translated to functional modulation in several previous studies [58–62] on small molecular inhibition of RNA targets, indicating that selectivity is more important than binding, to successfully inhibit downstream function of the RNA involved.

(b) **Exhaustive pocket prediction**: All possible non-overlapping binding sites present in apo RNA structures were extracted using the fpocket prediction program [63]. Overlap between two binding sites is defined based on the number of residues matching between the two sites. In this study, a cut-off of 90% match is used to resolve redundancy. This approach has also been previously followed by several studies on protein druggability prediction [32, 33, 39, 40, 64].

(c) **All-to-all flexible blind docking:** To emulate the negative binding sites defined based on NMR fragment screening assays and high-throughput screening in previous studies [29, 31, 65, 66], all unique RNA-binding ligands were docked to every unique RNA target in a blind fashion using the NLDock program [67]. Since in blind docking, the search space is set to the complete RNA structure, the same ligand can interact with multiple different binding sites in the structure. A maximum of 10 unique binding sites were sampled per RNA target through the flexible blind docking approach.

In summary, fpocket is designed for binding pocket prediction in proteins but can also detect geometric pockets in RNA, though it does not account for RNA-specific interactions. NLDock, used for all-to-all flexible blind docking, is specifically applicable to RNA-ligand docking. But it does not capture the druggable binding sites effectively (Supplementary Table S7). Consequently, the predicted binding pockets are likely less druggable and can be regarded as negative binding sites. The potential negative binding sites predicted by each of the above three approaches were combined. To resolve redundancy within the combined dataset, binding site overlap was calculated between all pairs of sites, and one copy of each redundant site was retained. 37, 6777 and 419028 sites were obtained from backbone motif search, exhaustive pocket prediction and all-to-all flexible blind docking, respectively. The total number of negative binding sites amounted to 425 842. Overlap analysis within the negative dataset, with a 70% residue overlap criterion to define high similarity, resulted in a dataset of 3 947 negative binding sites. A large number of binding sites were carefully omitted in this manner to make sure that the negative binding sites do not actually have high druggability. These sites were finally compared with the set of 819 redundant positive binding sites resulting in 92 non-redundant negative binding sites. The process of obtaining the 92 negative sites is illustrated in Supplementary Fig. S2 and the complete workflow is provided in Fig. 1.

Calculation of binding site features

A total of 69 structure-based features were curated based on literature survey for representation of binding sites to facilitate machine learning [39, 68–70]. They were computed for every binding site in the positive and negative datasets. These features were grouped into seven feature types: composition, pharmacophore, surface area, principal moments of inertia (PMI) (shape), roughness, torsions of binding site residues, and sugar puckering. The unique features calculated for each feature type are tabulated below (Table 1).

The pharmacophore-based features were calculated for every atom of the four standard nucleotides (A, C, G, U) using the PATTY atom-typing nomenclature [68] implemented in RDKit. The accessible surface area features were computed using the NACCESS program [71] and FreeSASA module [72] in *BioPython* [73]. The formulae for computing the PMI descriptors were obtained from the DrugPred_RNA study [39] (equations 1–4). The four grid spacing cut-offs (0.4, 0.8, 1.6, and 3.2) for roughness calculation were considered based on a previous study [69], where roughness was used as an indicator to predict protein-ligand binding sites from structures. The binding site torsions and sugar puckering parameters were obtained from the 3DNA web interface [70]. The other features were computed using custom python scripts based on the *BioPython* package.

$$Asphericity = \frac{0.5 * (PM3 - PM1)^2 + (PM3 - PM2)^2 + (PM3 - PM1)^2}{PM1^2 + PM2^2 + PM3^2} \quad (1)$$

$$Eccentricity = \frac{\sqrt{PM3^2 - PM1^2}}{PM3^2} \quad (2)$$

$$Spherocityindex = \frac{3 * PM1}{PM1 + PM2 + PM3} \quad (3)$$

$$Inertialshape factor = \frac{PM2}{PM1 * PM3} \quad (4)$$

where, PM1, PM2, and PM3 correspond to the three PMI of the point cloud of atoms belonging to the binding site residues.

Development and evaluation of the binding site druggability prediction model

The problem of binding site druggability prediction was cast as a binary classification problem, with the positive binding sites assigned to class 1 and negative binding sites assigned to class 0. The probabilities from the classification model were used to prioritize multiple binding sites present within the same RNA target [32, 33]. To finalize the features for predicting the druggability of a binding site, a forward feature selection approach was employed based on our previous study [74]. This approach enabled identification of feature combinations with no inter-feature correlation or multicollinearity, resulting in optimal model performance. Multiple different classification methods available in the *scikit-learn* python package [75] were tested, which include SVM, discriminant analyses (LDA and QDA), gaussian processes, tree-based methods, and boosting methods. Partial area under the ROC curve (pAUC) score and F1-score were used as metrics to quantify the model performance.

Stratified 10-fold cross validation was performed for evaluating the feature combination identified during model training. The best feature combination obtained was also used to predict the druggability of an external blind test dataset of binding sites. A detailed discussion about the external test dataset is provided in the 'Results and discussion' section. SHAPley analysis [76] was performed for model explainability, and the SHAP scores were extracted to rank the selected features according to their contribution to model predictions. Multiple case studies were performed to understand the significance of the features selected, potential of the model to generalize to apo structures and capture the effect of mutations on the druggability of the binding site.

Results and discussion

Statistics of the RNA-small molecule binding site datasets

A total of 819 binding sites were extracted from the unique chains of 399 RNA–ligand complex structures finalized for curation of the positive dataset, as explained in the Methods section. These binding sites were clustered based on Rfam annotations and manual classification to obtain 56

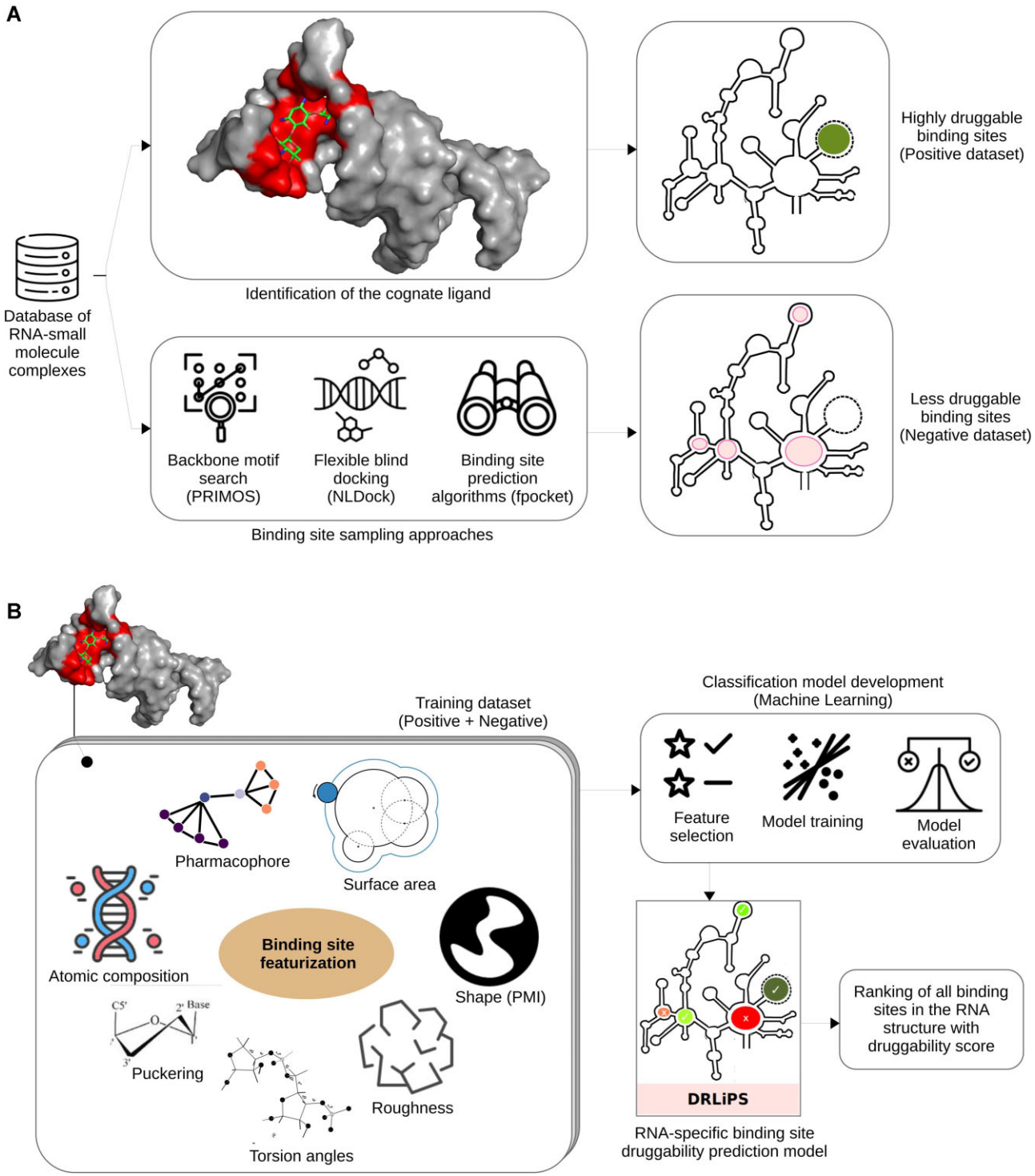


Figure 1. A schematic of the workflow followed for development of the DRLiPS model including (**A**) positive and negative dataset curation and (**B**) binding site featurization and model development..

non-redundant positive binding sites as detailed in the next section and Supplementary Table S2. To confirm the non-redundancy of the dataset after clustering, the binding site similarity was quantified for all binding site pairs using RM-score [77]. The resultant RMscore distributions clearly indicate a significant decrease in similarity between binding sites after clustering, which confirms the non-redundancy of the dataset (Supplementary Fig. S4).

The number of druggable binding sites per RNA in the positive dataset ranged from 1 to 5. The pocket size in terms of the number of binding site residues ranged between 4 and 30 in both the positive and negative datasets, with an average of 14 residues in the positive dataset and 10 residues in the negative dataset. The average volume of the binding sites was found to be 400.80 $Å^3$ in the positive dataset, and 398.66 $Å^3$ in the negative dataset. This shows that the negative dataset sam-

Table 1. Seven structure-based feature types used to represent the RNA-small molecule binding sites curated in this study

Feature type	Features	Feature count
Composition	A, C, G, U, Purines, Pyrimidines, Total atoms, Total heavy atoms, Total residues	9
Pharmacophore	HBA, HBD, Cation, Anion, Polar, Hydrophobic, Others	7
Surface area	Molecular surface area, Polar surface area, Non-polar surface area, Relative solvent accessible surface area (SASA), Relative polar SASA, Relative non-polar SASA, Pocket depth	7
Principal moments of inertia (Shape)	PMI1, PMI2, PMI3, NPR1, NPR2, Asphericity, Eccentricity, Spherocity index, Inertial shape factor	9
Roughness	R_0.4, R_0.8, R_1.6, R_3.2	4
Torsions of binding site residues	**Backbone torsions**: alpha, beta, gamma, delta, epsilon, zeta, e_z, chi, phase_angle, ssZp, Dp, splay	12
	Pseudo-torsions: eta, theta, eta', theta', eta", theta"	6
	Sugar torsions: v0, v1, v2, v3, v4, tm, P	7
Nucleotide type and sugar puckering	**Sugar type**: C3' endo, C2' endo, C3' exo, C2' exo **Pucker type**: C3' endo, C2' endo, C3' exo, C2' exo	8

pling strategy could successfully mimic the positive dataset in terms of the pocket size and volume, while other features such as surface area, composition, and pharmacophore can provide discrimination between the two sets.

RNA target families represented in the positive dataset

As discussed in the 'Materials and methods' section, the Infernal *cmscan* program [55] was used to identify the Rfam family assignments for the PDB structures present in the positive dataset. A total of 26 Rfam families could be mapped to the positive dataset, covering 225 (56.39%) of the 399 PDB structures. From the mapping exercise, it was noted that several popular Rfam families such as the 2′-deoxyguanosine riboswitch (RF01510), Pre-Q1 Class I riboswitch (RF00522), Guanidine Class II riboswitch (RF01068) etc., had only seed alignments deposited in Rfam, without any mapping to the PDB structures. In some cases, such as the recently discovered NAD + Class II riboswitch, Rfam families are yet to be defined. To account for these missing structures in the Rfam family assignments, the remaining 174 structures were manually assigned into one of 20 families, based on the deposited structure title and author-derived classification of the RNA target. Three of these 20 families (purine riboswitch, FMN riboswitch and SAM/SAH riboswitch) were found to have additional structures available in PDB, which are yet to be mapped in Rfam. Finally, all the 399 PDB structures could be assigned to one of 46 RNA target families provided in Supplementary Table S2. These 46 families were further subdivided to account for the presence of multiple different binding sites in the same RNA target, resulting in a total of 54 binding site families. For example, the prokaryotic ribosomal A-site family was sub-divided into five binding site families, to account for five different binding sites observed from ligand-bound PDB structures. The exact PDB structures corresponding to each RNA target family are provided in the Supplementary Table S1.

Details of the external blind test dataset

The external blind test dataset consists of empirical and modelled apo structures of RNA targets whose druggability has been verified. It consists of 15 positive and 70 negative binding pockets from the SARS-CoV-2 genomic elements and pre-miR-21 onco-miRNA. The SARS-CoV-2 genome has more than 60 structured elements [78] available in both the coding (ORF1ab) and non-coding regions (5′- and 3′-UTRs). However, the empirical druggability scores are available for only 11 prominent structured elements in the RNA genome [66], which are included in the test dataset (Supplementary Table S3). While a few of the prominent regions of the SARS-CoV-2 genome such as the ribosomal frameshifting element and stem-loop 2 (SL2) have crystal structures available in PDB, most of the genomic regions still remain unresolved. Hence, the tertiary structures of other RNA regions were taken from a genome-wide modelling study [79]. The positive binding pocket within the pre-miR-21 structure was identified from another study [80]. Additional negative binding pockets for each RNA target were sampled using the fpocket program. The similarity between pockets present in training dataset and external test dataset was also quantified using RMscore (Supplementary Fig. S7), which showed that ∼93% of the binding sites in the training set have less than 70% similarity to the test set. Further, three datasets were constructed from the test dataset with different RMscore cut-offs (0.7, 0.6, 0.5), which have 79, 62, and 25 data points, respectively, to evaluate the performance of the method.

Features chosen for the DRLiPS model

The SVM classifier with sigmoid kernel was chosen as the final model for druggability prediction. Normalization of the features between 0 and 1 using the *scikit-learn* MinMax normalization procedure yielded better results than using the unnormalized data for model development. The final set of six binding site features used to build the SVM model are: Hydrogen bonding (**HBA**) (No. of hydrogen bond acceptors), **C** (No. of Cytosines), **e_z** (epsilon-zeta backbone torsion angle), **alpha** (backbone torsion angle), **st_C2'_endo** (No. of sugars in the backbone with C2'-endo puckering), and **mol_sa** (accessible surface area). The distributions of these features in the positive and negative binding sites present in the training dataset are provided in Supplementary Fig. S8.

Interpretation of feature importance through SHAPley analysis and Principal Component analysis

SHAPley analysis is a popular approach to understand the contribution of each feature to model predictions, at both single data point-scale and complete model-scale. The analysis results in positive and negative scores for each feature, which are used to rank them in descending order of their importance to the model. For the SVM model with Sigmoid kernel, hereafter referred to as the DRLiPS model, SHAP analysis indicates that the feature categories contribute to druggability prediction in the following order: pharmacophore > composition > torsion > nucleotide type > surface area. Upon comparison of the distribution of the six features of the model be-

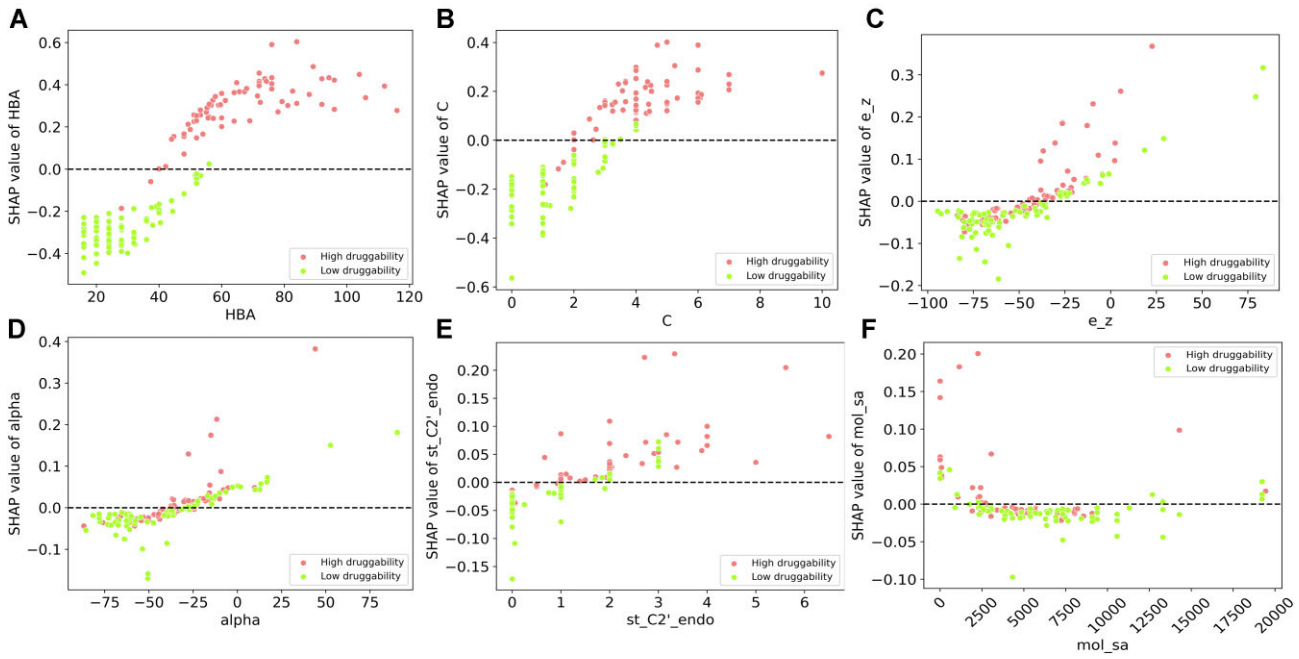


Figure 2. Distribution of SHAP values for the top 6 features with highly druggable and less druggable binding pockets. The distribution of SHAP values shows clear separation between the two types of binding pockets above and below zero SHAP value (dashed line).

tween the positive (highly druggable) and negative (less druggable) binding sites (Supplementary Fig. S3), HBA, C and mol_sa provide a clear discrimination in the feature distribution. Although the e_z, st_C2'_endo and alpha distributions overlap significantly between the two datasets, the positive distribution is multimodal in all three cases, compared to the unimodal negative distributions. Further, an ablation study involving the overlapping features showed that the model performance gradually improves with the addition of these features (Supplementary Table S11). In this way, all six features were found to be differentially distributed between the positive and negative datasets, justifying their selection for the final feature combination (Fig. 2).

The importance of these six features for predicting highly druggable binding sites in RNA is discussed below.

- **Hydrogen bond acceptor count (HBA):** HBA has been reported in several studies [39, 45, 56, 81, 82] to influence RNA-small molecule recognition significantly
- **C2'-endo sugar puckering (st_C2'_endo):** C2'-endo pucker is a sugar conformation in RNA structures, most commonly observed upon interaction with ligands. Nucleotides with C2'-endo sugar puckering (st_C2'_endo) have been previously shown to function as rate-limiting molecular switches in RNA conformation, due to their slow dynamics, which can potentially stabilize the RNA–ligand interactions [27, 48, 83]. Further, druggable (positive) binding sites show a higher prevalence of nucleotides with C2'-endo puckering than less druggable (negative) binding sites [83]. Interestingly, even apo RNA structures with nucleotides adopting the rare C2'-endo pucker were found to be involved in RNA–ligand interactions, with the puckered nucleotide in direct contact with the ligand [27].
- **Epsilon-Zeta (e_z) and alpha torsion angles:** Epsilon and zeta (e-z) are the most varying torsion angles of the RNA backbone, which enable backbone bending and stabilization in functionally important non-canonical RNA structures [84]. When modeling RNA–ligand interactions through docking, RNA backbone rotamers are extensively varied to capture the possible binding modes of RNA with the ligand [85]. Diverse conformational sampling by RNA backbone has also been implicated in slower association of ligand to binding pockets in RNA [86], highlighting the necessity to incorporate RNA backbone torsions during druggability prediction.
- **Cytosine count (C):** Cytosines present in RNA bulges have been shown to be crucial for recognition by small molecular probes [87]. Specifically, the number of cytosines and their pattern of distribution in RNA has been shown to impact the downstream effects of small molecule binding [87]. It is also notable that the presence of repeated cytosines in the binding site can encode disease-association of the RNA, as observed in repeat expansion diseases [88] and cancers [89].
- **Molecular surface area (mol_sa):** Surface area measures the solvent accessible region available in a given RNA structure for interaction with other biomolecules. Existing studies indicate that, surface area can be effectively used to quantify RNA conformational changes [90, 91], predict interaction hotspots [92], and predict the functional role of RNAs [93].

Further, we have examined the importance of selected features through Principal Component Analysis (PCA) with incremental addition of features based on the magnitude of variance captured by them, as inferred from the PCA loading matrix (Table 2). Based on a comparison of all pairs of principal components obtained from PCA, the first (PC1) and second (PC2) component pairs could provide the best separation between the highly druggable and less druggable binding sites (Fig. 3A). Next, the loading matrix was modified such that, only the top 2 features in PC1 and PC2 were retained (HBA and C in PC1; st_C2'_endo and e_z in PC2), and the rest of

Table 2. Loading matrix obtained from the PCA analysis on the training dataset. The maximum loading value observed for each feature is highlighted in bold

Feature/PC	PC1	PC2	PC3	PC4	PC5	PC6
mol_sa	−0.144707	0.164 159	**0.350 680**	−0.025131	0.018 847	0.019 526
C	**0.401 512**	−0.130619	0.113 708	0.065 983	0.106 761	−0.099410
st_C2'_endo	0.188 261	**0.321 204**	−0.062 192	−0.145227	−0.037328	−0.090249
e_z	0.014 425	**0.254 960**	−0.096534	0.086 991	0.178 147	0.066 096
alpha	−0.049456	**0.160 462**	0.000 006	0.227 291	−0.114201	−0.061037
HBA	**0.440 628**	0.045 362	0.041 285	0.016 334	−0.093795	0.126 542

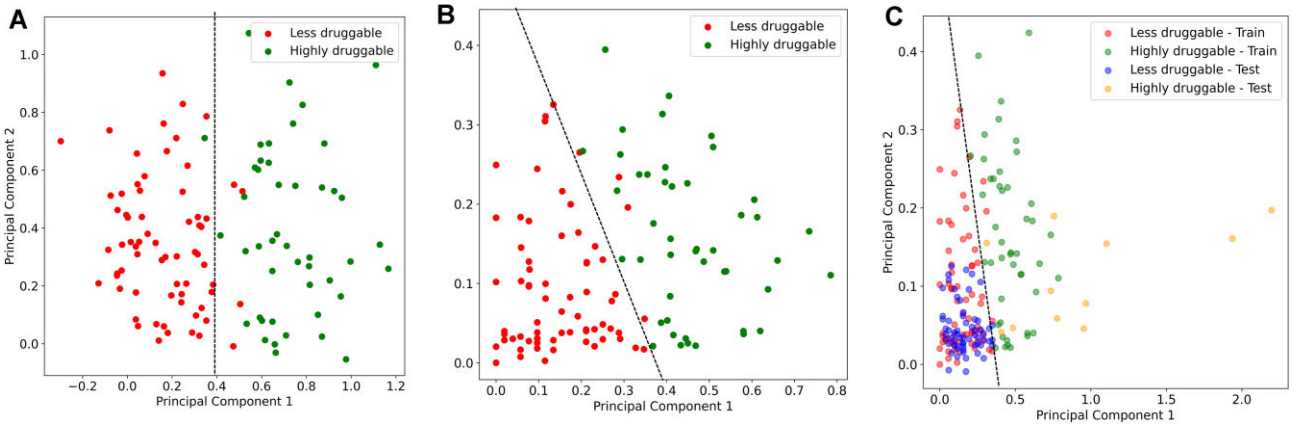


Figure 3. (**A**) Plot between PC1 and PC2 before modification of the loading matrix. The plot shows a clear separation between the binding site classes; (**B**) Plot after retaining only the loading values for HBA and C features in PC1, and for st_C2'_endo and e_z features in PC2. The plot still shows clear separation between the two classes for the training dataset; (**C**) Plot after overlaying the test data points after transformation with the modified loading matrix. The decision boundary between the two classes of binding sites is shown in each plot as a dashed line.

the feature loadings were set to zero. When the training and test datasets were transformed using the modified loading matrix, it was observed that the PC1 versus PC2 plot could still provide good separation between the two classes of binding sites (Fig. 3B). When the test data points were overlaid on the plot, a clear separation could still be observed between the two classes (Fig. 3C). From this analysis, it was observed that the features identified by the ML model development were also the features that capture maximum variance in the training dataset. The order of feature importance as inferred by their maximum PCA loading value was found to be: HBA (PC1) > C (PC1) > st_C2'_endo (PC2) > e_z (PC2) > alpha (PC2) > mol_sa (PC3). Except the positions of e_z and alpha features, the order of feature importance derived from PCA is in good agreement with the results from SHAP analysis. This indicates that PCA can also be used to explain the feature contribution in an ML model, with a much lower computational cost compared to SHAP analysis.

Comparison of model performance with existing methods

The model developed in this study can be compared to the DrugPred_RNA method [39] and RNACavityMiner method [40]. DrugPred_RNA model is trained using the NRDLD dataset of druggable proteins, and the descriptors identified are applied to RNA druggability prediction. The descriptors are computed using a superligand of the binding site of interest, which is the negative space occupied by a dataset of diverse ligands docked to the site. RNACavityMiner takes the RNA 3D structure in PDB format as the input and predicts all possible binding pockets using a custom spacefill algorithm. Five different ML models (MLP, XGBoost, Random forest, ExtraTrees and Logistic regresssion) are used to score each predicted binding pocket and their consensus score is also provided as output. For this comparison, the consensus score predicted by the program was extracted and predictions for different pocket overlap percentages were compared to the ground truth druggability classes (Supplementary Table S4). It must be noted that while DrugPred_RNA, RNACavityMiner and DRLiPS have been tested with the same external test dataset, the training dataset of DRLiPS contains additional RNA structures compared to both the methods, since it is based on the latest PDB release. Even with a relaxed cut-off of 50% overlap of residues between the predicted and observed pockets, the RNACavityMiner method could only achieve an F1-score = 0.15, Precision = 0.09, Specificity = 0.058, Recall = 0.45, and pAUC = 0.47. In case of DrugPred_RNA, the superligands were generated for each binding site present in the external test dataset, by docking a carefully chosen set of diverse ligands from their dataset. The best performance metrics achieved were: F1-score = 0.15, Precision = 0.13, Specificity = 0.95, Recall = 0.13, and pAUC = 0.54 (Supplementary Table S10). In comparison, DRLiPS achieves an F1-score = 0.70, Precision = 0.61, Specificity = 0.898, Recall = 0.73, and pAUC = 0.75 on the same external test dataset (Supplementary Table S3).

Further, three blind test sets were constructed by considering different RMscore cut-offs (0.7, 0.6, 0.5) and retaining only the dissimilar binding sites for the test dataset (Supplementary Section S2). The performance of DRLiPS model was evaluated on these three blind test sets and the results are provided in the Supplementary Table S6. On the 0.5 RMscore cut-off test set, the model was able to achieve

Table 3. Performance of various classification methods for RNA binding site druggability prediction. The partial AUC (pAUC) scores were obtained by truncating the ROC curve at 0.2 FPR (false positive rate), which translates to more than 80% specificity

Model	No. of features	Training pAUC	Training F1 score	Stratified 10-fold CV pAUC	Stratified 10-fold CV F1 score	Test set pAUC	Test set F1-score
SVM (Sigmoid kernel)	6	0.61	0.72	0.64	0.73	**0.75**	**0.70**
SVM (RBF kernel)	4	0.85	0.90	0.80	0.86	0.48	0.17
Gaussian Naïve Bayes	5	0.80	0.86	0.80	0.82	0.48	0.15
Decision Tree	5	1.0	1.0	0.77	0.78	0.53	0.23
AdaBoost	4	1.0	1.0	0.73	0.80	0.47	0.13
Random Forest	4	1.0	0.96	0.80	0.83	0.50	0.18
Gradient Boosting	3	1.0	1.0	0.80	0.79	0.56	0.41
XGBoost	3	0.93	0.94	0.82	0.82	0.63	0.43

a pAUC and F1-score of 0.95 and 0.94, respectively. The model is able to generalize well to binding sites with incrementally high dissimilarity to the training set as indicated by the performance metrics. This shows that DRLiPS can be compared to the DrugPred_RNA method [39] and RNACavityMiner method [40]. DrugPred_RNA model is trained using the NRDLD dataset of druggable proteins, and the descriptors identified are applied to RNA druggability prediction. The descriptors are computed using a superligand of the binding site of interest, which is the negative space occupied by a dataset of diverse ligands docked to the site. RNACavityMiner takes the RNA 3D structure in PDB format as the input and predicts all possible binding pockets using a custom spacefill algorithm. Five different ML models (MLP, XGBoost, Random forest, ExtraTrees and Logistic regresssion) are used to score each predicted binding pocket and their consensus score is also provided as output. For this comparison, the consensus score predicted by the program was extracted and predictions for different pocket overlap percentages were compared to the ground truth druggability classes (Supplementary Table S4). It must be noted that while DrugPred_RNA, RNACavityMiner and DRLiPS have been tested with the same external test dataset, the training dataset of DRLiPS contains additional RNA structures compared to both the methods, since it is based on the latest PDB release. In terms of the time taken for prediction on the external test dataset, DRLiPS is 120 times faster than RNACavityMiner, making it advantageous for integration into target prioritization pipelines for RNA-targeted drug discovery. A detailed discussion on the advantages of using DRLiPS over the existing methods is provided in Supplementary Section S3.

DRLiPS was also compared with multiple RNA-small molecule binding site prediction methods [94–96] on the same external test dataset. The results from this analysis are provided in Supplementary Table S7. The pAUC scores achieved by the three methods (MultiModRLBP = 0.51; RNAsite = 0.50; RLBind = 0.50) were much lower in comparison with DRLiPS (pAUC = 0.75), indicating that DRLiPS outperforms these methods on the test dataset. An additional dataset of 50 empirical RNA structures from PDB (released after 1 June 2024) were also curated for further validation of the model. Upon annotation of the RNA families present in this dataset, we found that 23 structures (46%) belong to families already present in the training dataset. Among the remaining structures, 18 (36%) were synthetic aptamer-small molecule complexes and 9 (17.6%) were non-coding RNA-small molecule complexes involved in human diseases. The performance of DRLiPS model on this new test set is provided in the Supplementary Table S8. The results show that the metrics (pAUC = 0.60; Top-3 accuracy = 0.82) obtained on the new test set are very similar to that obtained during training (pAUC = 0.61; Top-3 accuracy = 0.87). Hence, the features utilized in the DRLiPS model are able to generalize across widely varying RNA conformations from unseen RNA targets, including synthetic RNA aptamers and disease-associated non-coding RNAs.

Comparison with various machine learning methods

Table 3 summarizes the performance of the different classification models for RNA druggability prediction. The receiver-operating characteristic (ROC) curves for the SVM classifier are provided in Supplementary Fig. S10 to visualize the performance of the method in comparison with all other methods considered in Table 3. Also, the SVM model with Sigmoid kernel was found to outperform all the other machine learning methods on the external test dataset.

Generalizability of the model to apo RNA structures

Since the DRLiPS model was trained using only holo structures of RNA targets from PDB, this case study aims to showcase the potential of the model when apo structures are used as input. Through this analysis, the generalization capability of the model can be tested despite the use of features such as backbone torsion angles (e_z and alpha), which can vary extensively upon ligand-binding and result in marked conformational differences between the apo and holo states of the RNA [85, 86, 97]. Riboswitches have been shown to exhibit significant conformational change between their 'on' (ligand bound) and 'off' (unbound) states, which affects the ability of the switch to control gene expression downstream [98, 99]. A dataset of seven apo-holo riboswitch pairs was constructed from PDB and the druggable binding pockets were predicted using DRLiPS for both structures separately. Through binding pocket residue overlap analysis, matching pockets between apo and holo structure were identified and ranked. Spearman rank correlation coefficient was used to compare the ranking obtained for the overlapping binding pockets, to quantify the model's generalizability to apo structures (Table 4).

Based on the analysis, if the variation in backbone conformations between apo and holo states is more than 3 Å, the model's generalizability was found to be not very reliable, as in the case of TPP and PRPP riboswitches (Table 4). Interestingly, the backbone RMSD between apo and holo states and binding pocket rank correlation from DRLiPS were found to be highly inversely correlated ($r = -0.85$). This indicates that, as the backbone RMSD increases between apo and holo struc-

Table 4. A dataset of 7 apo-holo riboswitch structure pairs from PDB, used to test the generalizability of DRLiPS to apo structures. Backbone RMSD values were obtained from the RNA-align web server [106]. The table also shows the agreement in binding pocket ranking for the overlapping pockets between the structure pairs, based on druggability scores computed by DRLiPS web server

RNA target	Apo structure	Holo structure	Backbone RMSD from RNAlign program (Å)	Spearman's rank correlation between binding pocket ranks from DRLiPS
Adenine riboswitch	5E54	4TZX	2.57	0.642
THF riboswitch Class I	3SUY	3SUX	0.51	0.885
THF riboswitch Class II	7WIA	7WI9	0.77	0.8
TPP riboswitch	8F4O	2GDI	3.35	− 0.085
Pre-Q1 class I riboswitch	6VUH	6VUI	1.52	0.8
HCV IRES Domain IIa	1P5M	2KTZ	3.17	0.21
PRPP riboswitch	6DNR	6DLQ	2.04	0.785

tures, the agreement between binding site druggability scores from DRLiPS decreases and *vice versa*. It is also notable that, multiple recent articles [43, 100, 101] have highlighted the inability of experimental methods to capture the exact apo structure of RNA targets in solution, leading to highly similar apo and holo conformations in the PDB, which can be misleading. With availability of better representative apo state structures of RNA targets from experiments, the true generalizability of the method can be verified.

A potential solution to enable druggability prediction for highly dynamic RNA targets is to train the model with experimentally labelled apo conformations. Although several apo structures of small molecule-binding RNA are available in PDB, the binding site annotations are unknown from experimental studies for a majority of them. With the availability of a larger dataset of experimentally labelled apo-holo structure pairs with significant dynamics, the positive dataset for druggability prediction can be expanded to improve the model.

Ability of the DRLiPS druggability scores to capture the effect of single-point mutations on small molecule binding

Both active site and distal site mutations in RNAs have been reported to have significant effects on ligand binding, catalysis and subsequent regulation of cellular functions. To understand if the effect of such mutations can be captured through the druggability scores from DRLiPS, a dataset of 10 pairs of wild-type–mutant complexes with experimental structures and binding affinity (K_d) information were collected from the R-SIM database [102]. For these 10 pairs of complexes, the druggability scores for the cognate binding sites were predicted using the DRLiPS model and the scores were compared between the wild-type and mutant to understand the effects captured by the score (Table 5). For each unique RNA target in this dataset, the inferences arrived based on the score comparison are also provided below (Table 5) and in the Supplementary Table S5.

Based on the above observations, the DRLiPS druggability score has the ability to clearly distinguish the effects of single-point mutations on ligand binding. While most of the changes in predicted score are in agreement with the experimental observations, some conflicts were found to actually indicate the potential of the binding site to become more promiscuous towards other analogs of the cognate ligand, and thereby exhibit an increase in their druggability (SAM-VI riboswitch and Guanine riboswitch). In case of THF riboswitch, a marked decrease in the druggability score indicates the highly deleterious effect of the U25C mutation on the three-way junction binding site. While the relationship between selectivity and druggability of a binding site has not been explored yet for RNA targets, there are examples of proteins which exhibit high druggability and low selectivity (high promiscuity), indicating an inactive catalytic state [103–105]. The change in DRLiPS score between wild-type and mutants can be probed further in this regard to understand this aspect better, and arrive at a general relationship specific to selectivity of druggable RNA targets in future work.

To understand the sensitivity of the model to mutation effects, the mutated structure was obtained using three approaches: (i) directly from PDB (experimental), (ii) mutation module of Web3DNA [70] and (iii) mutagenesis module of PyMOL. We observed very low standard deviation in DRLiPS scores (Supplementary Table S9 and Supplementary Fig. S9) across the different mutant structures, which indicates that DRLiPS is a reliable model for prediction of the effect of binding site mutations on RNA druggability. At the same time, the average scores still confirm to the trend shown in Table 5, which is based on a single mutant residue conformation.

Web server development

A web server was developed to host the DRLiPS model for public usage. Given the 3D structure of an RNA target in PDB format, the DRLiPS web server can predict the druggability of binding pockets in two modes: 'Known site' mode and 'All sites' mode (Supplementary Fig. S5). The 'Known site' mode can be used during the following two scenarios: (i) the PDB structure has a bound ligand and the druggability needs to be quantified for the ligand-binding pocket, or (ii) the user needs to predict the druggability of a specific binding site, when the residues involved are known. Under these two scenarios, the prediction form can take either the PDB ligand ID (3-letter code), a comma-separated list of binding site residues, or a text file containing the binding site residues as input.

The 'All sites' mode can be used when the user does not have any information regarding the binding site, and would like to explore all possible binding sites in the structure. In this scenario, the input PDB structure is subject to pocket prediction using the fpocket program following which, the druggability of each binding pocket is predicted and reported in the Results. DRLiPS also includes the JSmol applet for visualization of multiple binding sites in the input structure in parallel (Supplementary Fig. S6). The web server was developed with HTML, CSS, JavaScript, Bootstrap, PHP, and Python packages. It is freely available at: https://web.iitm.ac.in/bioinfo2/DRLiPS/. The tutorials page of the web server includes exam-

Table 5. DRLiPS druggability scores predicted for 10 wild-type – mutant pairs of RNA–ligand complexes. The trend in binding affinity for the cognate ligand before and after mutation is also provided from the reference for each pair

PDB ID WT	PDB ID mutant	Target class[a]	Mutation	K_d WT (nM)	K_d mutant (nM)	K_d Trend[b]	WT score	Mutant score	Agreement with experimental studies
3mxh	3mum	c-di-GMP riboswitch	G20A	0.011	0.21 ± 0.07	D	0.64	0.63	In agreement
3mxh	3mur		C92U	0.011	15 ± 1	D	0.64	0.63	In agreement
6ck5	6ck4	PRPP riboswitch	G96A	2000 ± 300	1 600 000 ± 200 000	D	0.62	0.61	In agreement
3l3c	3g8t	glmS ribozyme	G33A	-	-	N	0.61	0.61	In agreement
4rzd	6xkn	PreQ1 Class III riboswitch	A52G	6.5 ± 0.5	4.0 ± 0.4	I	0.57	0.59	In agreement
4rzd	6xko		A84G	6.5 ± 0.5	27.2 ± 1.7	D	0.57	0.6	Not in agreement
2gis	2ygh	SAM-I riboswitch	G2nA (G19A)	540 ± 250	310 ± 60	I	0.59	0.57	Not in agreement
6las	6lax	SAM-VI riboswitch	U6C	330 ± 60	460 ± 20	D	0.54	0.6	In agreement
1y27	2b57	Guanine riboswitch	C74U	4 ± 3	-	L	0.51	0.56	In agreement
4lvv	3sd3	THF riboswitch	U25C	18 000 ± 1000	-	L	0.69	0.59	In agreement

[a]WT – Wild-type; GMP – Guanosine monophosphate; PRPP – Phosphoribosyl pyrophosphate; SAM – S-adenosyl methionine; THF – Tetrahydrofolate; G6P – Glucose-6-phosphate; FFO – 5-formyl tetrahydrofolate; K_d – dissociation constant.
[b]D – Decrease in binding affinity for cognate ligand; I – Increase in binding affinity for cognate ligand; L – Total loss of binding affinity for cognate ligand; N – No effect on binding affinity for cognate ligand

ples of usage for the two modes of prediction and the input file formats supported by the prediction form.

Potential applications for the DRLiPS model

DRLiPS can be utilized in RNA-targeted drug design pipelines in the following two scenarios:

- **Target identification**: Given an RNA target structure, DRLiPS can be used to find druggable binding sites of high confidence. The druggability score obtained from DRLiPS can be used to prioritize potential RNA targets for further experimental studies. Since the model can work with apo structures, modelled structures and mutated structures (with mutations present in binding site), it can be a valuable tool for target identification in early stages of drug design.
- **RNA optimization**: Since DRLiPS can predict the effect of mutations on the druggability of RNA-small molecule binding sites, targets with lower druggability scores can also be computationally optimized to improve their druggability. However, the mutations suggested by DRLiPS for RNA optimization should be subject to further experimental validation.

Conclusions

In this study, a model (DRLiPS) for prediction of potential druggable binding pockets in RNA structure was developed using known RNA-small molecule complex structures from PDB. The features selected for model development could capture the polar nature of RNA-small molecule binding pockets, dynamic nature of the RNA backbone during ligand association, and the favorable torsions and puckering signifying successful ligand-binding events. Although the model was trained only with holo RNA structures, the model could generalize well to apo RNA structures when the conformational variation was within 3 Å. Consequently, DRLiPS outperforms the existing prediction method in the domain on an external test dataset comprising both apo and modelled RNA structures. Further, the model could capture the effect of single point mutations on druggability, with promiscuous binding sites predicted to retain high druggability after mutation. However, the model could not capture the effect of far-site mutations and could not distinguish loss of selectivity post-mutation, which will be addressed in our future studies. The DRLiPS model has been hosted as a web server for public use and is freely available at: https://web.iitm.ac.in/bioinfo2/DRLiPS/.

Acknowledgements

The authors thank the members of Protein Bioinformatics lab and the Computational Structural Biology team from TCS Research for their valuable suggestions and inputs for model development and evaluation.

CRediT author contributions: **Sowmya Ramaswamy Krishnan**: Data Curation, Methodology, Validation, Formal analysis, Software, Writing – Original Draft, Writing – Review & Editing, Visualization. **Arijit Roy**: Conceptualization, Validation, Resources, Writing – Review & Editing, Supervision. **Limsoon Wong**: Conceptualization, Methodology, Validation, Investigation, Writing – Review & Editing, Supervision. **M. Michael Gromiha**: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – Review & Editing, Supervision.

Supplementary data

Supplementary data are available at NAR Online.

Conflict of interest

S.R.K. and A.R. are employed by Tata Consultancy Services Limited.

Funding

This work is partially supported by the Indian Council of Medical Research (ICMR) to MMG (Grant number: EMDR/SG/15/2024-01-01272).

Data availability

All datasets and source codes associated with this work are available at https://github.com/Sowmya-R-Krishnan/DRLiPS and https://doi.org/10.6084/m9.figshare.28582004.

References

1. Warner KD, Hajdin CE, Weeks KM. Principles for targeting RNA with drug-like small molecules. *Nat Rev Drug Discov* 2018;**17**:547–58. https://doi.org/10.1038/nrd.2018.93
2. Liu J, Dou X, Chen C *et al.* N^6-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription. *Science* 2020;**367**:580–6. https://doi.org/10.1126/science.aay6018
3. Jame-Chenarboo F, Ng HH, Macdonald D *et al.* High-throughput analysis reveals miRNA upregulating α-2,6-sialic acid through direct miRNA-mRNA interactions. *ACS Cent Sci* 2022;**8**:1527–36. https://doi.org/10.1021/acscentsci.2c00748
4. Li Y, Tong Y, Liu J *et al.* The role of MicroRNA in DNA damage response. *Front Genet* 2022;**13**:850038. https://doi.org/10.3389/fgene.2022.850038
5. Xi X, Li T, Huang Y *et al.* RNA biomarkers: frontier of precision medicine for cancer. *Noncoding RNA* 2017;**3**:9.
6. Toden S, Goel A. Non-coding RNAs as liquid biopsy biomarkers in cancer. *Br J Cancer* 2022;**126**:351–60. https://doi.org/10.1038/s41416-021-01672-8
7. Pardi N, Hogan MJ, Porter FW *et al.* mRNA vaccines - a new era in vaccinology. *Nat Rev Drug Discov* 2018;**17**:261–79. https://doi.org/10.1038/nrd.2017.243
8. Zhu Y, Zhu L, Wang X *et al.* RNA-based therapeutics: an overview and prospectus. *Cell Death Dis* 2022;**13**:644. https://doi.org/10.1038/s41419-022-05075-2
9. Aagaard L, Rossi JJ. RNAi therapeutics: principles, prospects and challenges. *Adv Drug Deliv Rev* 2007;**59**:75–86. https://doi.org/10.1016/j.addr.2007.03.005
10. Traber GM, Yu A. RNAi based therapeutics and novel RNA bioengineering technologies. *J Pharmacol Exp Ther* 2023;**384**:133–54. https://doi.org/10.1124/jpet.122.001234
11. Mullard A. FDA approves RNA-targeting small molecule. *Nat Rev Drug Discov* 2020;**19**:659.
12. Mattick JS, Amaral PP, Carninci P *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 2023;**24**:430–47. https://doi.org/10.1038/s41580-022-00566-8
13. Andrews RJ, Baber L, Moss WN. RNAStructuromeDB: a genome-wide database for RNA structural inference. *Sci Rep* 2017;**7**:17269. https://doi.org/10.1038/s41598-017-17510-y
14. Hughes JP, Rees S, Kalindjian SB *et al.* Principles of early drug discovery. *Br J Pharmacol* 2011;**162**:1239–49. https://doi.org/10.1111/j.1476-5381.2010.01127.x
15. Schenone M, Dančík V, Wagner BK *et al.* Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* 2013;**9**:232–40. https://doi.org/10.1038/nchembio.1199
16. Raies A, Tulodziecka E, Stainer J *et al.* DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets. *Commun Biol* 2022;**5**:1291. https://doi.org/10.1038/s42003-022-04245-4
17. Emmerich CH, Gamboa LM, Hofmann MCJ *et al.* Improving target assessment in biomedical research: the GOT-IT recommendations. *Nat Rev Drug Discov* 2021;**20**:64–81. https://doi.org/10.1038/s41573-020-0087-3
18. Zhou Y, Zhang Y, Zhao D *et al.* TTD: therapeutic Target Database describing target druggability information. *Nucleic Acids Res* 2024;**52**:D1465–77. https://doi.org/10.1093/nar/gkad751
19. Barril X. Druggability predictions: methods, limitations, and applications. *WIREs Comput Mol Sci* 2013;**3**:327–38. https://doi.org/10.1002/wcms.1134
20. Finan C, Gaulton A, Kruger FA *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* 2017;**9**:eaag1166. https://doi.org/10.1126/scitranslmed.aag1166
21. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. *Bioinformatics* 2009;**25**:451–7. https://doi.org/10.1093/bioinformatics/btp002
22. Bull SC, Doig AJ. Properties of protein drug target classes. *PLoS One* 2015;**10**:e0117955. https://doi.org/10.1371/journal.pone.0117955
23. Gianni D, Farrow S. Functional genomics for target identification. *SLAS Discov* 2020;**25**:531–4. https://doi.org/10.1177/2472555220927692
24. Haley B, Roudnicky F. Functional genomics for cancer drug target discovery. *Cancer Cell* 2020;**38**:31–43. https://doi.org/10.1016/j.ccell.2020.04.006
25. Moffat JG, Vincent F, Lee JA *et al.* Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;**16**:531–43. https://doi.org/10.1038/nrd.2017.111
26. Kubota K, Funabashi M, Ogura Y. Target deconvolution from phenotype-based drug discovery by using chemical proteomics approaches. *Biochim Biophys Acta (BBA) - Proteins Proteom* 2019;**1867**:22–7. https://doi.org/10.1016/j.bbapap.2018.08.002
27. Kligun E, Mandel-Gutfreund Y. Conformational readout of RNA by small ligands. *RNA Biol* 2013;**10**:981–9. https://doi.org/10.4161/rna.24682
28. Zeller MJ, Favorov O, Li K *et al.* SHAPE-enabled fragment-based ligand discovery for RNA. *Proc Natl Acad Sci USA* 2022;**119**:e2122660119. https://doi.org/10.1073/pnas.2122660119
29. Hajduk PJ, Huth JR, Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* 2005;**48**:2518–25. https://doi.org/10.1021/jm049131r
30. Cheng AC, Coleman RG, Smyth KT *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**:71–5. https://doi.org/10.1038/nbt1273
31. Gupta A, Gupta AK, Seshadri K. Structural models in the assessment of protein druggability based on HTS data. *J Comput Aided Mol Des* 2009;**23**:583–92. https://doi.org/10.1007/s10822-009-9279-y
32. Halgren TA. Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* 2009;**49**:377–89. https://doi.org/10.1021/ci800324m
33. Sheridan RP, Maiorov VN, Holloway MK *et al.* Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J Chem Inf Model* 2010;**50**:2029–40. https://doi.org/10.1021/ci100312t
34. Krasowski A, Muthas D, Sarkar A *et al.* DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *J Chem Inf Model* 2011;**51**:2829–42. https://doi.org/10.1021/ci200266d
35. Perola E, Herman L, Weiss J. Development of a rule-based method for the assessment of protein druggability. *J Chem Inf Model* 2012;**52**:1027–38. https://doi.org/10.1021/ci200613b
36. Volkamer A, Kuhn D, Rippmann F *et al.* DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* 2012;**28**:2074–5. https://doi.org/10.1093/bioinformatics/bts310
37. Liu T, Altman RB. Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS Comput Biol* 2011;**7**:e1002326. https://doi.org/10.1371/journal.pcbi.1002326
38. Hussein HA, Borrel A, Geneix C *et al.* PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res* 2015;**43**:W436–42. https://doi.org/10.1093/nar/gkv462
39. Rekand IH, Brenk R. DrugPred_RNA-A tool for structure-based druggability predictions for RNA binding sites. *J Chem Inf Model* 2021;**61**:4068–81. https://doi.org/10.1021/acs.jcim.1c00155

40. Xie J, Frank AT. Mining for ligandable cavities in RNA. *ACS Med Chem Lett* 2021;**12**:928–34. https://doi.org/10.1021/acsmedchemlett.1c00068
41. Cunningham M, Pins D, Dezső Z *et al.* PINNED: identifying characteristics of druggable human proteins using an interpretable neural network. *J Cheminform* 2023;**15**:64. https://doi.org/10.1186/s13321-023-00735-7
42. Zhou Y, Chen S. Advances in machine-learning approaches to RNA-targeted drug design. *Artif Intell Chem* 2024;**2**:100053. https://doi.org/10.1016/j.aichem.2024.100053
43. Lee HK, Lee Y, Fan L *et al.* Crystal structure of Escherichia coli thiamine pyrophosphate-sensing riboswitch in the apo state. *Structure* 2023;**31**:848–59. https://doi.org/10.1016/j.str.2023.05.003
44. Schroeder KT, Daldrop P, Lilley DMJ. RNA tertiary interactions in a riboswitch stabilize the structure of a kink turn. *Structure* 2011;**19**:1233–40. https://doi.org/10.1016/j.str.2011.07.003
45. Menichelli E, Lam BJ, Wang Y *et al.* Discovery of small molecules that target a tertiary-structured RNA. *Proc Natl Acad Sci USA* 2022;**119**:e2213117119. https://doi.org/10.1073/pnas.2213117119
46. Burley SK, Bhikadiya C, Bi C *et al.* RCSB Protein Data bank: tools for visualizing and understanding biological macromolecules in 3D. *Protein Sci* 2022;**31**:e4482. https://doi.org/10.1002/pro.4482
47. Katz AM, Tolokh IS, Pabit SA *et al.* Spermine condenses DNA, but not RNA duplexes. *Biophys J* 2017;**112**:22–30. https://doi.org/10.1016/j.bpj.2016.11.018
48. David-Eden H, Mankin AS, Mandel-Gutfreund Y. Structural signatures of antibiotic binding sites on the ribosome. *Nucleic Acids Res* 2010;**38**:5982–94. https://doi.org/10.1093/nar/gkq411
49. Lin J, Zhou D, Steitz TA *et al.* Ribosome-targeting antibiotics: modes of action, mechanisms of resistance, and implications for drug design. *Annu Rev Biochem* 2018;**7**:451–78. https://doi.org/10.1146/annurev-biochem-062917-011942
50. Knox C, Wilson M, Klinger CM *et al.* DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res* 2024;**52**:D1265–75. https://doi.org/10.1093/nar/gkad976
51. Zdrazil B, Felix E, Hunter F *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* 2024;**52**:D1180–92. https://doi.org/10.1093/nar/gkad1004
52. Desaphy J, Bret G, Rognan D *et al.* sc-PDB: a 3D-database of ligandable binding sites–10 years on. *Nucleic Acids Res* 2015;**43**:D399–404. https://doi.org/10.1093/nar/gku928
53. Peng X, Liao W, Lin X *et al.* Crystal structures of the NAD+-II riboswitch reveal two distinct ligand-binding pockets. *Nucleic Acids Res* 2023;**51**:2904–14. https://doi.org/10.1093/nar/gkad102
54. Kalvari I, Nawrocki EP, Ontiveros-Palacios N *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* 2021;**49**:D192–200. https://doi.org/10.1093/nar/gkaa1047
55. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5. https://doi.org/10.1093/bioinformatics/btt509
56. Padroni G, Patwardhan NN, Schapira M *et al.* Systematic analysis of the interactions driving small molecule-RNA recognition. *RSC Med Chem* 2020;**11**:802–13. https://doi.org/10.1039/D0MD00167H
57. Duarte CM, Wadley LM, Pyle AM. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res* 2003;**31**:4755–61. https://doi.org/10.1093/nar/gkg682
58. Hendrix M, Priestley ES, Joyce GF *et al.* Direct observation of aminoglycoside-RNA interactions by surface plasmon resonance. *J Am Chem Soc* 1997;**119**:3641–8. https://doi.org/10.1021/ja964290o
59. Trausch JJ, Batey RT. A disconnect between high-affinity binding and efficient regulation by antifolates and purines in the tetrahydrofolate riboswitch. *Chem Biol* 2014;**21**:205–16. https://doi.org/10.1016/j.chembiol.2013.11.012
60. Kelly ML, Chu C, Shi H *et al.* Understanding the characteristics of nonspecific binding of drug-like compounds to canonical stem-loop RNAs and their implications for functional cellular assays. *RNA* 2021;**27**:12–26. https://doi.org/10.1261/rna.076257.120
61. Martin WJ, Grandi P, Marcia M. Screening strategies for identifying RNA- and ribonucleoprotein-targeted compounds. *Trends Pharmacol Sci* 2021;**42**:758–71. https://doi.org/10.1016/j.tips.2021.06.001
62. Tong Y, Lee Y, Liu X *et al.* Programming inactive RNA-binding small molecules into bioactive degraders. *Nature* 2023;**618**:169–79. https://doi.org/10.1038/s41586-023-06091-8
63. Guilloux VL, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf* 2009;**10**:168. https://doi.org/10.1186/1471-2105-10-168
64. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem* 2010;**53**:5858–67. https://doi.org/10.1021/jm100574m
65. Binas O, de Jesus V, Landgraf T *et al.* ^{19}F NMR-based fragment screening for 14 different biologically active RNAs and 10 DNA and protein counter-screens. *ChemBioChem* 2021;**22**:423–33. https://doi.org/10.1002/cbic.202000476
66. Sreeramulu S, Richter C, Berg H *et al.* Exploring the druggability of conserved RNA regulatory elements in the SARS-CoV-2 genome. *Angew Chem Int Ed* 2021;**60**:19191–200. https://doi.org/10.1002/anie.202103693
67. Feng Y, Zhang K, Wu Q *et al.* NLDock: a fast nucleic acid-ligand docking algorithm for modeling RNA/DNA-ligand complexes. *J Chem Inf Model* 2021;**61**:4771–82. https://doi.org/10.1021/acs.jcim.1c00341
68. Bush BL, Sheridan RP. PATTY: a programmable atom type and language for automatic classification of atoms in molecular databases. *J Chem Inf Comput Sci* 1993;**33**:756–62. https://doi.org/10.1021/ci00015a015
69. Todoroff N, Kunze J, Schreuder H *et al.* Fractal dimensions of macromolecular structures. *Mol Inf* 2014;**33**:588–96. https://doi.org/10.1002/minf.201400090
70. Li S, Olson WK, Lu X. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res* 2019;**47**:W26–34. https://doi.org/10.1093/nar/gkz394
71. Hubbard SJ, Thornton JM. *'NACCESS', Computer Program*. London: Department of Biochemistry and Molecular Biology, University College, 1993.
72. Mitternacht S. FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res* 2016;**5**:189. https://doi.org/10.12688/f1000research.7931.1
73. Cock PJA, Antao T, Chang JT *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3. https://doi.org/10.1093/bioinformatics/btp163
74. Krishnan SR, Roy A, Gromiha MM. Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. *Brief Bioinform* 2024;**25**:bbae002. https://doi.org/10.1093/bib/bbae002
75. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
76. Lundberg S, Lee SA. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, 4768–77. https://dl.acm.org/doi/10.5555/3295222.3295230
77. Zheng J, Xie J, Hong X *et al.* RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genom* 2019;**20**:276. https://doi.org/10.1186/s12864-019-5631-3
78. Huston NC, Wan H, Strine MS *et al.* Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel

regulatory motifs and mechanisms. *Mol Cell* 2021;**81**:584–98. https://doi.org/10.1016/j.molcel.2020.12.041
79. Manfredonia I, Nithin C, Ponce-Salvatierra A *et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res* 2020;**48**:12436–52. https://doi.org/10.1093/nar/gkaa1053
80. Costales MG, Matsumoto Y, Velagapudi SP *et al.* Small molecule targeted recruitment of a nuclease to RNA. *J Am Chem Soc* 2018;**140**:6741–4. https://doi.org/10.1021/jacs.8b01233
81. Sengupta RN, Herschlag D. Enhancement of RNA/ligand association kinetics via an electrostatic anchor. *Biochemistry* 2019;**58**:2760–8. https://doi.org/10.1021/acs.biochem.9b00231
82. Falese JP, Donlic A, Hargrove AE. Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chem Soc Rev* 2021;**50**:2224–43. https://doi.org/10.1039/D0CS01261K
83. Gherghe CM, Mortimer SA, Krahn JM *et al.* Slow conformational dynamics at C2'-endo nucleotides in RNA. *J Am Chem Soc* 2008;**130**:8884–5. https://doi.org/10.1021/ja802691e
84. Zgarbová M, Luque FJ, Sponer J *et al.* Toward improved description of DNA backbone: revisiting Epsilon and Zeta Torsion force field parameters. *J Chem Theory Comput* 2013;**9**:2339–54. https://doi.org/10.1021/ct400154j
85. Stefaniak F, Chudyk EI, Bodkin M *et al.* Modeling of ribonucleic acid–ligand interactions. *WIREs Comput Mol Sci* 2015;**5**:425–39. https://doi.org/10.1002/wcms.1226
86. Gleitsman KR, Sengupta RN, Herschlag D. Slow molecular recognition by RNA. *RNA* 2017;**23**:1745–53. https://doi.org/10.1261/rna.062026.117
87. Das B, Murata A, Nakatani K. A small-molecule fluorescence probe ANP77 for sensing RNA internal loop of C, U and A/CC motifs and their binding molecules. *Nucleic Acids Res* 2021;**49**:8462–70. https://doi.org/10.1093/nar/gkab650
88. Koshy BT, Zoghbi HY. The CAG/polyglutamine tract diseases: gene products and molecular pathogenesis. *Brain Pathol* 1997;**7**:927–42. https://doi.org/10.1111/j.1750-3639.1997.tb00894.x
89. Maulik A, Bandopadhyay D, Singh M. A cytosine-patch sequence motif identified in the conserved region of lincRNA-p21 interacts with the KH3 domain of hnRNPK. *Curr Res Struct Biol* 2023;**5**:100099. https://doi.org/10.1016/j.crstbi.2023.100099
90. Rouskin S, Zubradt M, Washietl S *et al.* Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 2014;**505**:701–5. https://doi.org/10.1038/nature12894
91. Mukherjee S, Bahadur RP. An account of solvent accessibility in protein-RNA recognition. *Sci Rep* 2018;**8**:10546. https://doi.org/10.1038/s41598-018-28373-2
92. Barik A, Nithin C, Karampudi NB *et al.* Probing binding hot spots at protein-RNA recognition sites. *Nucleic Acids Res* 2016;**44**:e9. https://doi.org/10.1093/nar/gkv876
93. Solayman M, Litfin T, Singh J *et al.* Probing RNA structures and functions by solvent accessibility: an overview from experimental and computational perspectives. *Brief Bioinform* 2022;**23**:bbac112. https://doi.org/10.1093/bib/bbac112
94. Su H, Peng Z, Yang J. Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics* 2021;**37**:36–42. https://doi.org/10.1093/bioinformatics/btaa1092
95. Wang K, Zhou R, Wu Y *et al.* RLBind: a deep learning method to predict RNA-ligand binding sites. *Brief Bioinform* 2023;**24**:bbac486. https://doi.org/10.1093/bib/bbac486
96. Wang J, Quan L, Jin Z *et al.* MultiModRLBP: a deep learning approach for multi-modal RNA-small molecule ligand binding sites prediction. *IEEE J Biomed Health Inform.* 2024;**28**:4995–5006. https://doi.org/10.1109/JBHI.2024.3400521
97. Noeske J, Buck J, Fürtig B *et al.* Interplay of 'induced fit' and preorganization in the ligand induced folding of the aptamer domain of the guanine binding riboswitch. *Nucleic Acids Res* 2007;**35**:572–83. https://doi.org/10.1093/nar/gkl1094
98. Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol* 2011;**3**:a003533. https://doi.org/10.1101/cshperspect.a003533
99. Juru AU, Patwardhan NN, Hargrove AE. Understanding the contributions of conformational changes, thermodynamics, and kinetics of RNA-small molecule interactions. *ACS Chem Biol* 2019;**14**:824–38. https://doi.org/10.1021/acschembio.8b00945
100. Stagno JR, Liu Y, Bhandari YR *et al.* Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. *Nature* 2017;**541**:242–6. https://doi.org/10.1038/nature20599
101. Wu MT, D'Souza V. Alternate RNA structures. *Cold Spring Harb Perspect Biol* 2020;**12**:a032425. https://doi.org/10.1101/cshperspect.a032425
102. Krishnan SR, Roy A, Gromiha MM. R-SIM: a database of binding affinities for RNA-small molecule interactions. *J Mol Biol* 2023;**435**:167914. https://doi.org/10.1016/j.jmb.2022.167914
103. Cheng AC. Predicting selectivity and druggability in drug discovery. *Annu Rep Comput Chem* 2008;**4**:23–37.
104. Jacobs MD, Caron PR, Hare BJ. Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: structure of lck/imatinib complex. *Proteins* 2008;**70**:1451–60. https://doi.org/10.1002/prot.21633
105. Cerisier N, Petitjean M, Regad L *et al.* High impact: the role of promiscuous binding sites in polypharmacology. *Molecules* 2019;**24**:2529. https://doi.org/10.3390/molecules24142529
106. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* 2019;**35**:4459–61. https://doi.org/10.1093/bioinformatics/btz282

Received: October 2, 2024. **Revised:** February 16, 2025. **Editorial Decision:** March 11, 2025. **Accepted:** March 14, 2025