JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

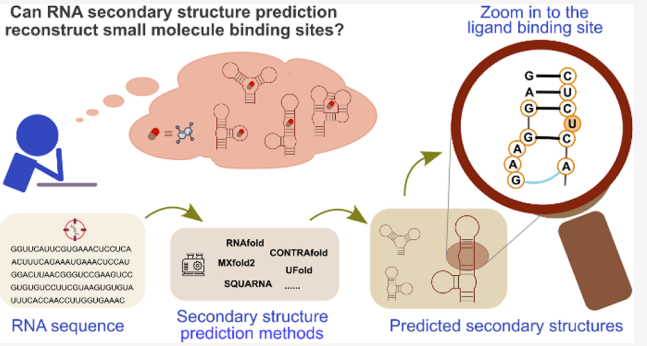
pubs.acs.org/jcim
Article

Exploring Secondary Structure Predictions for RNA-Targeted Drug Discovery: Power and Challenges

Zhengyue Zhang,* Gaia Dolcetti, Christian Tyrchan, Leonardo De Maria, Giovanni Bussi, and Werngard Czechtizky

Cite This: https://doi.org/10.1021/acs.jcim.6c00108
Read Online

ABSTRACT: RNAs are increasingly recognized as promising drug targets, as both coding and noncoding RNAs act as key regulators in disease-related biological processes. However, a significant gap persists between the number of known RNA sequences and the solved RNA structures, posing a major bottleneck for RNA-targeted drug discovery. RNA secondary structure prediction offers the potential to facilitate the identification of druggable sites in novel RNA sequences by rapidly predicting base pairing patterns. In this study, we benchmarked widely used RNA secondary structure prediction tools against a newly curated dataset of ligand-bound RNA structures. We found that most tools achieve reasonable accuracy for RNAs with short sequences and simple motifs, but their performance declines for longer RNAs and those containing pseudoknots. Notably, prediction accuracy is reduced within ligand binding sites, where noncanonical base pairs and complex secondary structure elements are prevalent yet consistently unrecognized by the tools. Consequently, RNA ligand binding sites are poorly reconstructed by secondary structure predictions. This work provides the first comprehensive assessment of RNA secondary structure prediction for ligand-bound RNAs and demonstrates the challenges for integrating these methods into RNA-targeted drug discovery pipelines.



INTRODUCTION

RNA has attracted increasing research interest due to its diverse cellular roles.[1−5] It has been widely recognized that small molecules can target RNAs as therapeutics since RNAs are involved in different biological processes, including molecular mechanisms underlying human diseases.[6,7] Several categories of RNAs and RNA-associated complexes have been identified as therapeutic targets for small molecules, such as rRNAs, riboswitches, and splicing modulators.[8−11] Notably, the FDA approval of Risdiplam—the first RNA-targeting small molecule drug, which regulates pre-mRNA splicing activity[12,13]—has spurred significant interest from the pharmaceutical industry in developing RNA-targeted therapeutics.[8,14−21] Approximately 70% of the genome encodes RNAs, compared to 1.5% for proteins, and many of these transcripts are linked to key biological processes and diseases.[1,18] Therefore, it is prospective to expand the space of druggable targets by taking RNAs into account.

Structure-based drug discovery depends on detailed target structures to identify binding pockets and guide the design of compounds. While the Protein Data Bank (PDB) contains extensive structural information for proteins, it includes only about 1,000 RNA-small molecule complexes.[22,23] This scarcity of RNA structural data represents a major challenge for RNA-targeted drug design. When presented with a target RNA sequence of unknown structure, RNA structure prediction offers a rapid means to gain primary structural insights. Among the available approaches, RNA secondary structure predictions emerged as a valuable computational tool in this context.[24] The secondary structure of RNA describes its base pairing patterns within RNAs (A-U, G-C, G-U, and noncanonical base pairs), determining the molecule's overall folding. Accurately predicting secondary structure from sequence can locate complex secondary structure elements, such as bulge loops, internal loops, hairpin loops, etc., on the sequence, which often form the compound binding pockets.[8,24,25] Thus, the RNA secondary structure prediction can serve as a critical component of RNA-targeted drug discovery pipelines.

RNA secondary structure prediction methodologies have evolved over the past 40 years, employing a variety of strategies. Classical prediction tools, such as RNAfold[26,27] and

Received: January 12, 2026
Revised: March 17, 2026
Accepted: March 18, 2026

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

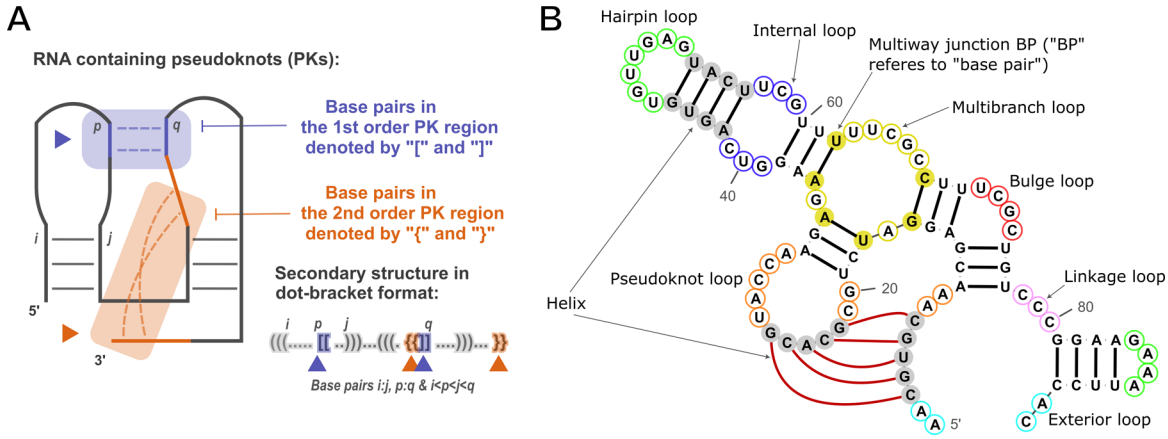


Figure 1. (A) Schematic representation of pseudoknots (PKs). An RNA structure contains a PK when it has at least two base pairs, b_i:b_j and b_p:b_q, that satisfy the index condition i < p < j < q, representing non-nested base pairing. The "orders of PKs" refer to the number of distinct bracket types (other than "()") required to annotate PK base pairs in dot-bracket format. (B) Secondary structure elements that are derived from canonical base pairs.

RNAstructure,[28,29] rely on thermodynamic parameters of base pairs from Turner's nearest-neighbor model and use Zuker's dynamic programming algorithm to search for the optimal secondary structure.[24,30,31] With the advent of artificial intelligence, machine learning (ML)-based prediction methods, such as MXfold2 and UFold, further expanded the field, demonstrating improved prediction accuracy.[32−36] More recently, language model (LM)-based approaches have captured secondary structure information from RNA sequences and achieved competitive performance compared with classical and ML-based methods.[37−39] Over time, advancements have also enabled the prediction of complex RNA features, including pseudoknots (PKs) and noncanonical base pairs. The enriched details given by PK and noncanonical base pair predictions improved reconstruction of ligand binding sites. PKs are characterized by non-nested base pairing (Figure 1A)[40,41] and are often found in functionally important regions, likely contributing to ligand binding site formation.[29,42,43] Accurate PK prediction is thereby essential for drug discovery purposes, and various secondary structure prediction methods and PK-specialized tools are available for PK-containing RNAs, such as RNAstructure-ProbKnot,[28,44] iPKnot,[45,46] SQUARNA,[47] and CParty.[48] On the other hand, noncanonical base pairs are also pivotal in stabilizing complex RNA structures and in forming ligand interaction surfaces.[49,50] To date, only a few ML-based methods are capable of noncanonical base pair prediction.[33,51,52] However, the accuracy of noncanonical base pair predictions remains limited. Beyond traditional single-sequence-based approaches, multiple sequence alignment-based methods leveraging evolutionary information for RNA secondary structure prediction have been developed and can yield improved prediction accuracy.[53,54] Additionally, specialized tools such as RNAprobe and RNAthor integrate experimental data, particularly from chemical probing, to achieve improved accuracy of secondary structure predictions.[55−60] However, both alignment-based and experimentally driven prediction approaches are constrained by the availability of homologous sequences or experimental data. As a result, traditional single-sequence-based tools continue to serve as the primary choice for most RNA secondary structure predictions.

RNA secondary structure prediction tools have been benchmarked across diverse datasets, presenting acceptable full-length prediction accuracy on relatively short RNAs (less than 1000 nucleotides).[33,51,53,61] However, it remains unclear whether these tools can be reliably integrated into an RNA-targeted drug discovery pipeline to accurately reconstruct compound binding pockets and thereby enable downstream applications (e.g., virtual screening for hit finding). Critically, evaluation must also focus on performance within complex secondary structure elements at binding sites, such as helices, hairpin loops, bulge loops, and so on (Figure 1B). Furthermore, RNA structures may significantly alter upon ligand binding; for instance, riboswitches adopt distinct secondary structures depending on ligand presence.[62−64] Secondary structure predictions were not developed to account for ligand binding. Whether secondary structure prediction tools capture RNA topologies corresponding to ligand-bound or -unbound states remains an important question for drug discovery researchers.

In this work, we benchmarked a range of RNA secondary structure prediction tools, focusing specifically on ligand-bound RNAs with 3D structures. For this purpose, we constructed a dataset that includes entries from HARIBOSS,[23] the RNA-ligand complexes collected by Nithin et al.,[65] and newly released complexes in PDB.[22] The high-resolution data enable a detailed assessment of the prediction performance at ligand binding sites. We benchmarked tools representing a variety of approaches, including those based on thermodynamic parameters as well as ML methods. The tools include: RNAfold,[26,27] RNAstructure-Fold,[28,29] RNAstructure-MaxExpect,[28,66] RNAstructure-ProbKnot,[28,44] LinearFold,[67] iPKnot,[45,46] SPOT-RNA2,[52,53] CentroidFold-CONTRAfold,[68,69] CentroidFold-McCaskill,[68,70] CONTRAfold,[69] KnotFold,[51] REDfold,[71] E2Efold,[72] UFold, UFold_nc (UFold with noncanonical base pairing prediction activated),[33] MXfold2,[32,36] and SQUARNA.[47] We did not include LM-based approaches in our benchmarking because they have not reached a broad user base, do not outperform classical or ML-based methods, and remain in early development. To evaluate the contribution of RNA secondary structure prediction to structure-based drug discovery pipelines, the methods listed above are sufficient. Our analyses provide a comprehensive assessment of prediction performance for both overall RNA secondary structures and ligand binding sites. The evaluation of secondary structure element prediction within ligand binding sites offers a complementary view of tool performance in drug

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

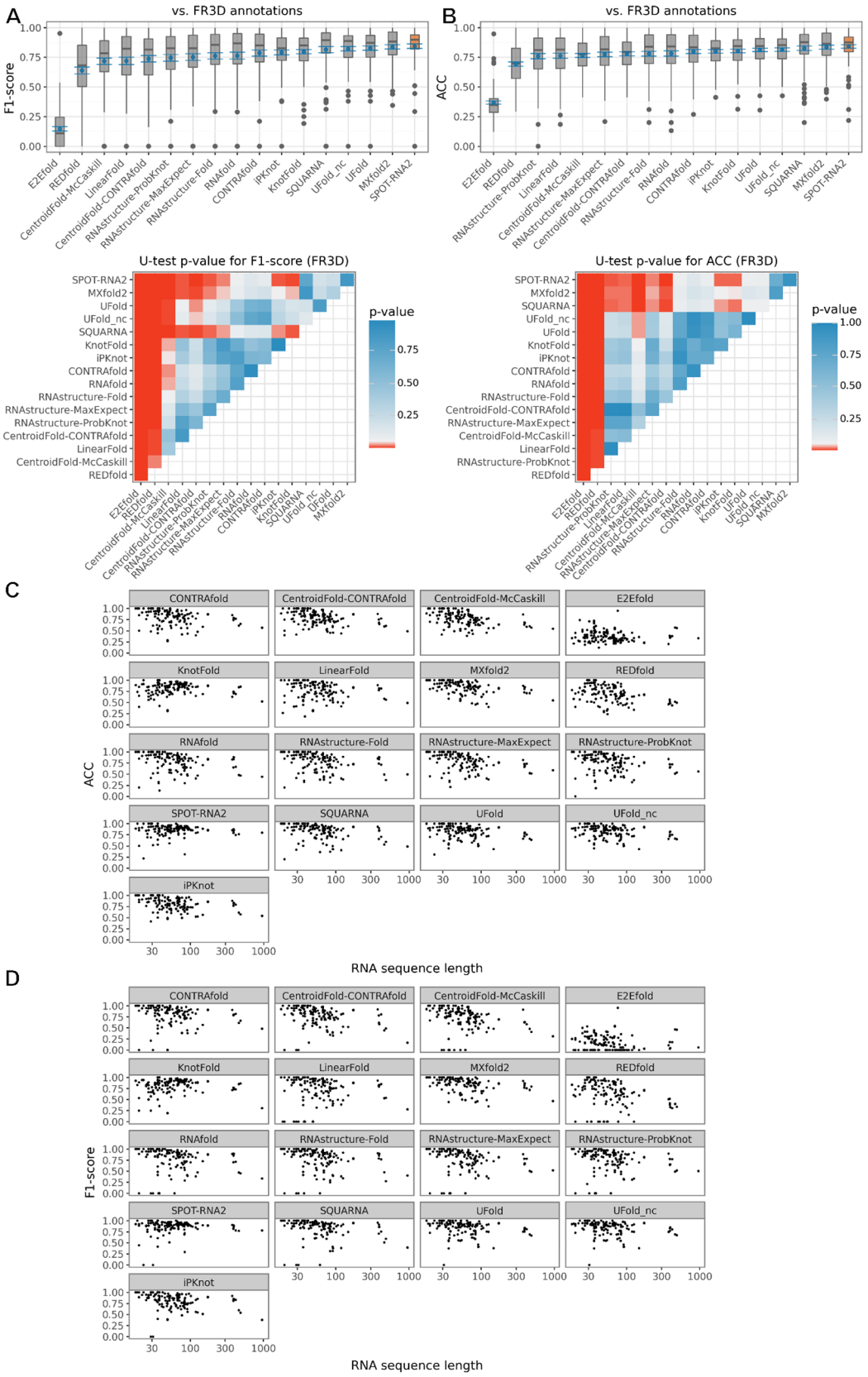


Figure 2. F1-score (A) and ACC value (B) distributions on the ligand-bound RNAs given by all prediction tools (top). Average values are indicated by dots, with the blue error bars representing the standard error of the mean. The tools are reordered by the averaged F1-score and ACC values, respectively. U-test was conducted to evaluate the statistical significance of the F1-score and ACC value differences, which are shown by *p*-values (bottom). Distributions of F1-score (C) and ACC value (D) versus sequence lengths of predicted RNAs.

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

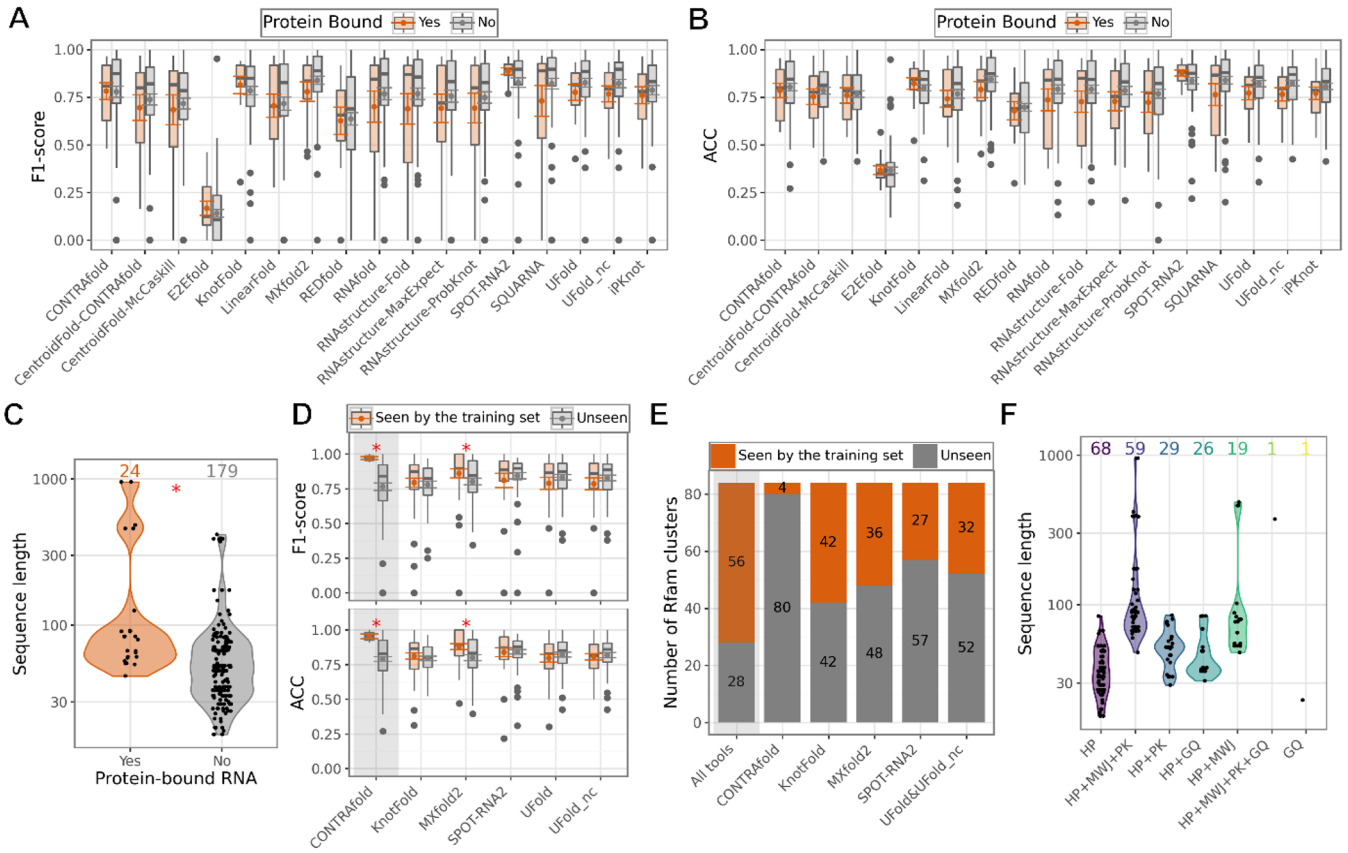


Figure 3. Comparisons of F1-score (A) and ACC value (B) distributions on protein-bound and unbound RNAs are given by all prediction tools. Average values are indicated by dots, with error bars representing the standard error of the mean. No significant difference was observed for either F1-score or ACC between predictions for protein-bound and nonprotein-bound RNAs. (C) Correlation between RNA length and protein binding; the difference in sequence length distributions is significant. (D) Comparisons of F1-score and ACC value distributions on data being or not being part of the training set of each machine learning-based prediction tool. Average values are indicated by dots, with error bars representing the standard error of the mean. The significant differences are labeled by the red asterisk. (E) The numbers of data being and not being part of the training set of each machine learning-based prediction tool. (F) Correlation between RNA length and RNAs containing different combinations of structure motifs. The numbers of RNAs with different combinations of motifs are listed at the top of the figure.

discovery. Additionally, the availability of 3D RNA-ligand complex structures enabled us to evaluate noncanonical base pair prediction for tools with this functionality (e.g., KnotFold,[51] UFold_nc,[33] SPOT-RNA2[52,53]).

Our results indicate that most prediction tools achieved satisfactory overall accuracy on ligand-bound RNAs. However, accurate base pairing prediction at ligand binding sites remains challenging for all of the tools. The situation is further complicated by the limited recognition of structure elements and the concentrated noncanonical base pairs in these regions. Consequently, although RNA secondary structure prediction can offer primary structure insights for target RNAs, extensive downstream computational and experimental efforts are needed to validate and refine the predicted models. Interestingly, although secondary structure prediction tools are not designed to model ligand-bound conformations, their outputs may provide hints regarding the ligand-induced structural dynamics of RNAs.

RESULTS

To evaluate the capability of RNA secondary structure prediction in identifying RNA ligand binding sites, we curated a dataset comprising 203 RNA-ligand complexes, organized into 84 clusters according to Rfam families. For benchmarking purposes, RNA sequences and their experimentally determined secondary structures were extracted. We evaluated 17 prediction tools, representing diverse methodological categories, by applying each tool to every RNA in the dataset. The predicted base pairing patterns were first analyzed at the level of overall RNA topology, followed by a focused assessment of the predictions in RNA ligand binding sites. Details for dataset curation, selected prediction tools, and benchmarking metrics are provided in the Materials and Methods.

Benchmarking of the Secondary Structure Prediction Tools at Full-Length Level

We evaluated overall RNA topology prediction against the FR3D-annotated reference secondary structure using F1-score and accuracy (ACC) (see Materials and Methods for details). The F1-score is the harmonic mean of precision and sensitivity for base pair prediction, providing a single metric that balances both. In contrast, ACC takes the correct identification of unpaired bases into account and thus gives equal importance to predictions of paired and unpaired bases, including various loop elements. Overall, almost all evaluated prediction tools show comparable F1-scores and ACC, with average values generally ranging from 0.70 to 0.90, except for E2Efold and REDfold, which perform significantly worse than the others (Figure 2A,B). Among all tools, SPOT-RNA2 appears to be the best-performing one in terms of metrics; however, its

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

performance does not differ statistically from those of some other competitive tools, such as UFold_nc and MXfold2. We also assessed the F1-score and ACC values with secondary structure annotations generated by x3dna-dssr. The results are consistent with the analysis above, as only a small proportion of base pairs were annotated differently between the two annotation methods (Figure S1). The multiple sequence alignment-based tool SPOT-RNA2 shows clear advantages in accurately predicting RNA secondary structures in our ligand-bound RNA dataset. However, since SPOT-RNA2 relies on evolutionary information that is not available for artificial RNA aptamers, its prediction accuracy on these RNAs is lower than that of MXfold2 and SQUARNA (Figure S2).

Several structural features contribute to the folding complexity of RNAs and challenge secondary structure prediction tools. The distributions of both F1-score and ACC values decrease noticeably for all evaluated tools as RNA length increases (Figure 2C,D). Similarly, the prediction accuracies for protein-bound RNAs are lower compared to those for nonprotein-bound RNAs (Figure 3A,B), while those protein-bound RNAs are generally longer than their counterparts (Figure 3C). Regarding the machine learning-based prediction tools, we compared the prediction accuracies for RNAs redundant with their training sets or not. Surprisingly, only CONTRAfold and MXfold2 show significantly better performance on the sequences present in their training sets (Figure 3D). Note that only four clusters in our dataset are redundant with the CONTRAfold training set (Figure 3E). The evaluated tools also display varied prediction performances across different combinations of RNA motifs. For most tools, RNAs with a hairpin (HP)-only topology are more likely to be accurately predicted. In contrast, RNAs containing multiway-junction (MWJ), pseudoknot (PK), or G-quadruplex (GQ) motifs present greater challenges for secondary structure predictions than HP-only RNAs (Figure S3). Although long RNAs are more likely to contain non-HP motifs (Figure 3F), this tendency persists when considering only RNAs with lengths between 50 and 100 nucleotides (Figure S3). It should be noted that the distribution of motif combinations in the dataset is highly imbalanced; especially, there is only one RNA with GQ-only topology and one containing all motif types. The averaged values of the metrics of different prediction tools discussed above are summarized in Table S2.

Pseudoknot Predictions Remain Challenging

Among the 84 RNA clusters in our dataset, 38 contain PKs (refer to 88 RNAs). Accurate identification of base pairs in PK regions and correct assignment of PK order—evaluated by strict PK base pair hit rates (see Materials and Methods)—are fundamental for capturing the correct RNA topology. Several of the evaluated prediction tools—namely, RNAstructure-ProbKnot, iPKnot, SPOT-RNA2, KnotFold, REDfold, E2Efold, UFold, UFold_nc, and SQUARNA—were designed to consider PK prediction. The strict PK base pair hit rates for these tools vary widely, from 0 to 1, across RNAs containing PKs (Figure 4). Based on average hit rate values, SPOT-RNA2, KnotFold, UFold, and UFold_nc are the top performers in recognizing base pairs within PK regions. However, even the best-performing tool, SPOT-RNA2, achieves an average hit rate of less than 0.5, indicating that accurate PK recognition remains a challenge for RNA secondary structure prediction. Although many tools were not designed to predict PKs, certain methods enable hierarchical identification of PK base pairs

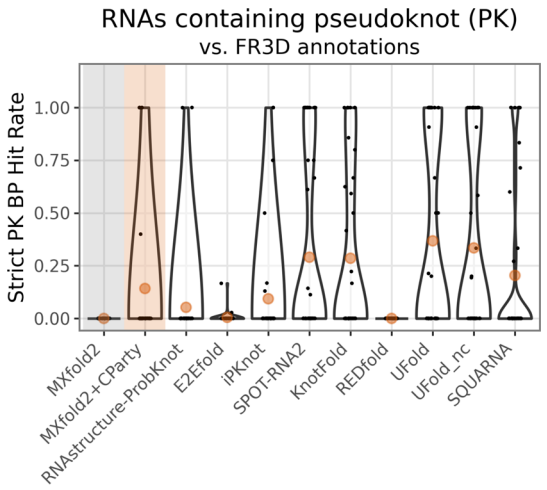


Figure 4. Distribution of strict PK base pair (BP) hit rates on ligand-bound RNAs as predicted by different tools. Orange dots represent the average hit rates among all clusters, and black dots are the hit rates for individual RNA clusters. Prediction tools incapable of pseudoknot prediction have a strict PK base pair hit rate of zero; therefore, their results are not shown.

from non-PK outputs. We applied CParty to predict PKs using the secondary structures predicted by MXfold2, which was the best-performing tool among those not specifically designed for PK prediction.[48] However, the strict PK hit rates achieved by MXfold2+CParty did not outperform those of the dedicated PK prediction tools (Figure 4). We also noticed that all evaluated prediction tools may predict base pairs in PK regions without assigning them to the correct PK orders (Figure S4). In other words, although base pairs in PK regions are identified, the tools missed the base pairs outside the PK regions that are necessary to form the non-nested base pairs. Consequently, the overall PK topology is either not predicted at all or assigned incorrectly.

Less Prediction Accuracy on RNA Ligand Binding Pockets

Accurate prediction of base pairing patterns in ligand-contacted nucleotides is crucial for subsequent 3D modeling of ligand binding pockets, docking, and any structure-based design (e.g., free energy perturbation). The ligand-contacted nucleotides contain atoms within 6 Å from the ligands, according to HARIBOSS criteria.[23] When comparing ACC values between ligand-contacted and nonligand-contacted regions, we found that all prediction tools except E2Efold have generally lower accuracy in ligand-contacted regions. These differences are generally small, and only CONTRAfold, KnotFold, SPOT-RNA2, SQUARNA, UFold, and UFold_nc exhibit significantly lower accuracy in these regions (Figure 5A). The ligand-contacted regions typically comprise a small number of nucleotides, which may introduce bias into the results (Figure 5B). Nonetheless, our findings suggest an increased difficulty in predicting the base pairing patterns within ligand-contacted regions.

Poor Recognitions of Complex Local Secondary Structure Elements at Ligand Binding Sites

RNAs frequently adopt various secondary structure elements—referred to as "structure elements" below—within ligand binding sites to shape binding pockets. According to Turner's nearest-neighbor model,[30,31] structure elements, based on canonical base pairs, are classified as helices, bulge

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

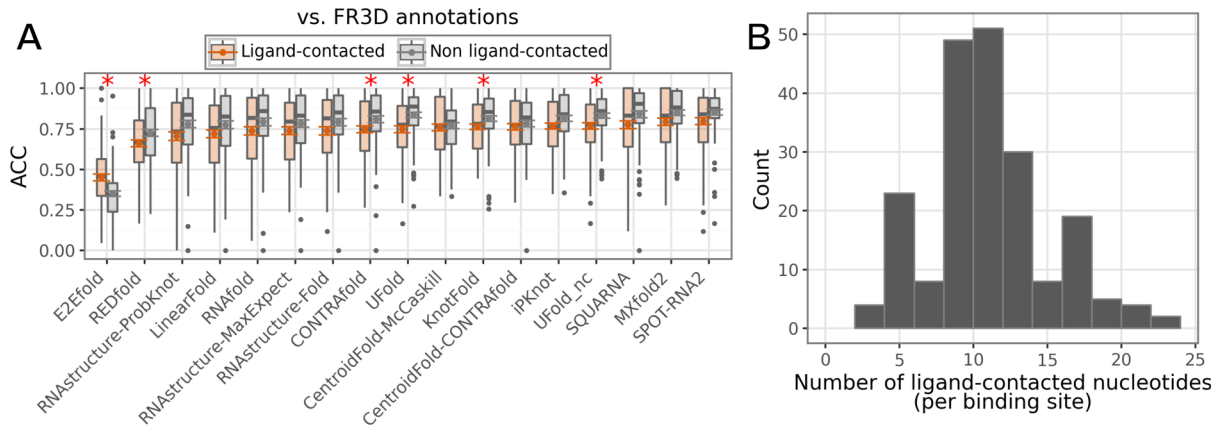


Figure 5. (A) Comparison of ACC values between ligand-contacted and nonligand-contacted regions across all prediction tools. Average values are indicated by dots, with error bars representing the standard error of the mean. Significant differences in performance between the two regions are indicated by asterisks. (B) Distribution of the number of ligand-contacted nucleotides per binding site.

loops, internal loops, multibranch loops, multiway junction BP ("BP" refers to "base pair") and exterior loops. We further expand this set by including pseudoknot loops and linkage loops, which are not covered by Turner's model. Pseudoknot loops connect helical stems comprising non-nested base pairs and capture the structure complexity of pseudoknots. Linkage loops are loops not encompassed by any previously defined element and are not enclosed by base pairs or other loops (Figure 1B). Notably, "pseudoknot" structure element in Turner's model is labeled as helices, since they do not differ from helices at the 3D level.

We annotated structure elements across RNAs in our dataset and quantified the occurrence probability of each element within ligand-contacted regions. Helices are the most common structural elements in those regions. Among the loop elements, internal loops are the most prevalent, with an occurrence rate exceeding 50%. Intriguingly, pseudoknot loops are the most frequent loop type in ligand-contacted regions of PK-containing RNAs, highlighting the need for a specific definition for this loop class (Figure 6A). In contrast, multibranch loops and multiway junction BPs are less likely to occur within ligand-contacted regions among RNAs that contain MWJ.

Accurate recognition of these elements is, therefore, a critical prerequisite for identifying RNA ligand binding sites. We examined whether prediction tools can correctly reconstruct secondary structure elements beyond merely achieving accurate prediction of base pairing patterns in those regions. For all clusters, we counted the number of ligand-contacted nucleotides that were correctly assigned to each structure element by the prediction tools (Figure 6B). The top-performing tools, SPOT-RNA2, UFold, and UFold_nc, exhibited high accuracy for the helix element, correctly assigning nearly 400 out of approximately 500 ligand-contacted nucleotides to helices across the dataset. Most tools also accurately reconstruct bulge loops and hairpin loops within these regions. Despite internal loops being common within ligand-contacted regions, only MXfold2 correctly assigned more than half of the relevant bases to this category. Recognition of minor elements, such as exterior loops and multibranch loops, was generally poor. For tools capable of pseudoknot prediction, the accurate assignment of pseudoknot loops within ligand-contacted regions remained limited, consistent with the generally modest performance of pseudoknot prediction. Collectively, the evaluated secondary structure prediction tools demonstrate inadequate accuracy in identifying local secondary structure elements within ligand-contacted regions.

Low Reliability in Noncanonical Base Pair Prediction

Historically, the field of RNA secondary structure prediction focused only on canonical base pairs, A-U, C-G, and the wobble pair G-U. However, noncanonical base pairs (e.g., A-A, G-A, and pairs that involve non-cis-Watson−Crick interactions) are common in RNAs and play significant roles in RNA folding.[73−77] Among the ligand-bound RNAs in our dataset, only 7 RNAs (from 4 different clusters) lack noncanonical base pairs. More importantly, the proportion of noncanonical base pairs among all pairing interactions is generally higher in ligand-contacted regions than in nonligand-contacted regions (Figure 7A).

Noncanonical base pairing interactions occur as single pairs and as triplexes, G-quartets, and other types of multiplexes at ligand-interacting surfaces. We identified all bases involved in noncanonical base pairing and annotated relevant bases as additional structure elements: GQ-loop, G-quartet, single noncanonical base pair, triplex, and multiplex (Figure 7B). Interestingly, both the single noncanonical base pair and triplex exhibit occurrence rates above 50% within ligand-contacted regions, underscoring a substantial contribution of non-canonical base pairing—in multiple forms—to RNA ligand binding pocket formation. Therefore, accurate prediction of noncanonical base pairs is as important as prediction of canonical base pairs for reconstructing RNA ligand binding sites.[50,78,79]

We evaluated several tools designed to predict noncanonical base pairs, assessing their ability to identify these interactions. All prediction tools capable of such analysis can report noncanonical base pairs in various base−base combinations (e.g., A-A, A-G) (Figure S5). However, unlike the high prediction accuracy observed for canonical base pairs, these tools rarely succeeded in identifying noncanonical base pairs (Figures 7D and S6). For instance, although UFold_nc achieved the highest hit rate for noncanonical base pairs, its rates were below 40% for most RNAs, as indicated by the upper quartile. Hit rates for other tools dropped to 30% or less. Notably, tools incapable of predicting noncanonical base pairs still showed some hits, as several predicted base pairs in

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Figure 6. (A) Possibility of observing different secondary structure elements within ligand-contacted regions. The occurrence possibilities of pseudoknot or multiway junction-associated structure elements are calculated only for RNAs containing pseudoknots or multiway junctions, respectively. (B) Number of bases correctly assigned to each secondary structure element compared with the total number for that element in ligand-contacted regions (overlapped). Results for pseudoknot loops are only shown for tools capable of pseudoknot prediction.

canonical combinations correspond to non-cis-Watson−Crick A-U, C-G, and G-U pairs in the reference structures.

Within ligand-contacted regions, the hit rates of non-canonical base pairs are even lower. The top-performing tools, UFold_nc and SPOT-RNA2, showed decreased accuracy compared with their predictions on the full RNA sequences, with most RNAs having hit rates below 25%. Other tools designed for noncanonical base pair prediction identified fewer than 5% of such base pairs in most RNAs. Many noncanonical base pairs are found as parts of triplexes, G-quartets, or other multiplexes in binding pockets, yet most tools predicted only a single pair and missed all these structure elements involving over two bases (Figure 7E). Generally, secondary structure prediction tools restrict each base to pair with one other base; some newer tools, such as UFold_nc, allow multiplex prediction. We found that UFold_nc predicted 20 triplexes across the dataset, but none matched those in the reference structures. As a result, the limited prediction accuracy on noncanonical base pairs highlights a major challenge for *de novo* modeling of RNA ligand binding sites with current secondary structure prediction methods.

WC-Edge Base-Ligand Contact

Within ligand-binding pockets, the WC-edge of RNA nucleobases is often involved in hydrogen-bonding interactions with ligands, which are critical for ligand recognition (Figure 8A).[80,81] If these bases are incorrectly assigned as paired by the prediction tools, then the interaction surface of the ligand-binding pocket will be disrupted. Among the collected 203 ligand-bound RNA structures, more than half of the PDB entries (114) present one or more nucleobases that interact

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

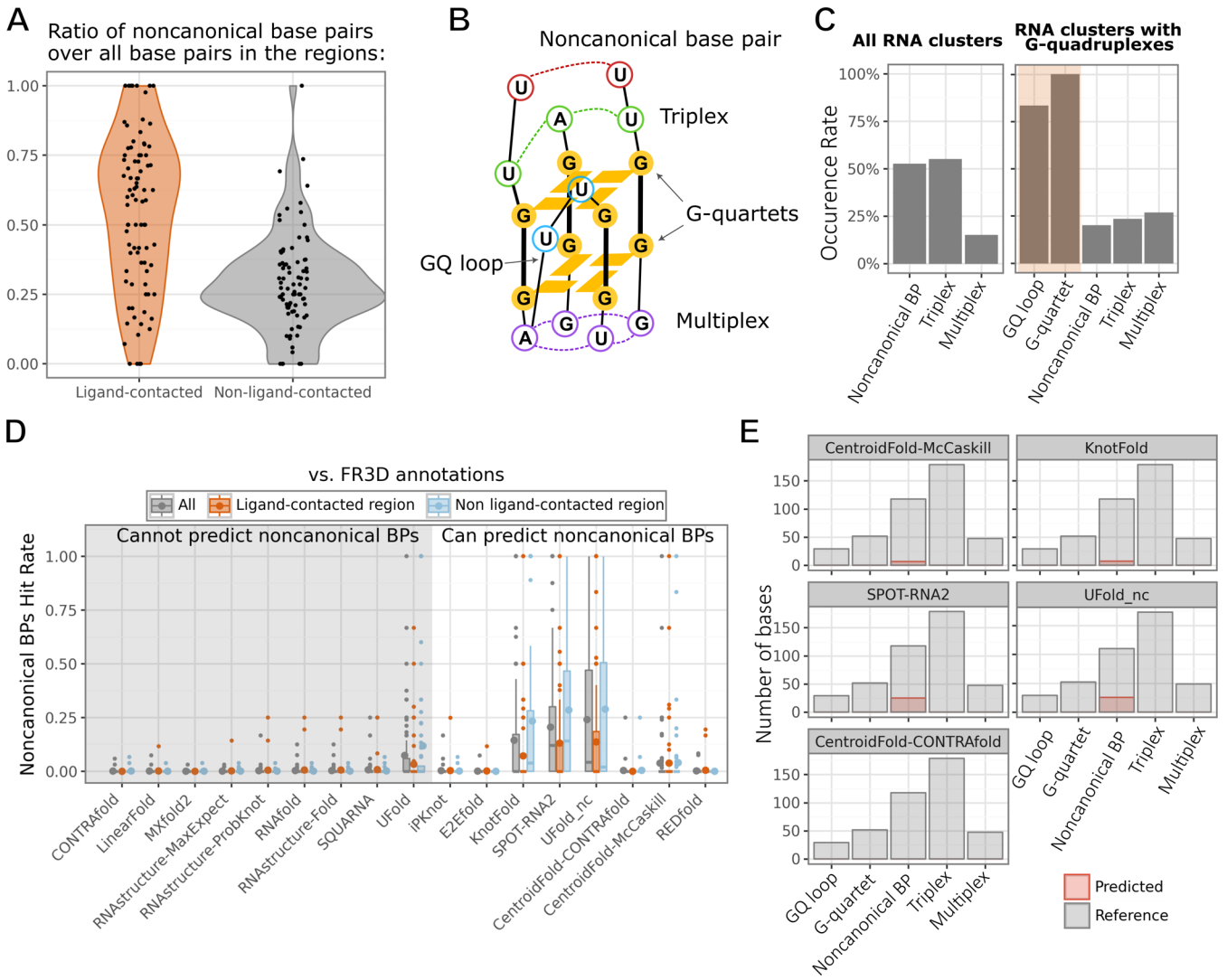


Figure 7. (A) Proportions of noncanonical base pairs among all base pairs in ligand-contacted and nonligand-contacted regions; each dot represents an individual RNA. (B) Scheme of structure elements defined by noncanonical base pairs. (C) Possibility of observing noncanonical base pair-associated elements within ligand-contacted regions. G-quadruplex (GQ)-relevant elements were analyzed only among RNAs containing GQs. (D) Distributions of hit rates for noncanonical base pairs in three regions: the full length, ligand-contacted, and nonligand-contacted regions. (E) Number of bases correctly assigned to each noncanonical base pair-associated element versus the total number for that element in ligand-contacted regions (overlapped). Only machine learning-based tools are included.

with the bound ligand through their WC-edges (Figure 8B). This highlights the importance of correctly predicting these bases as unpaired in order to reconstruct ligand-binding pockets. Therefore, we evaluated whether secondary structure prediction tools can correctly identify WC-edge ligand-contacted bases as unpaired.

We measured the rate at which different prediction tools incorrectly assigned WC-edge ligand-contacted bases as paired; see Figure 8C. Such bases are likely to be predicted as unpaired, as indicated by the averaged buried rate of around 25% across all tools. Therefore, the WC-edges of these bases are likely to be exposed in predicted secondary structures for ligand contacting. However, it is important to note that H-bonding interactions between ligands and nucleobases critically influence binding affinity and specificity; therefore, exposing the WC-edges of all WC-edge ligand-contacted bases within an RNA ligand binding site is critical. The evaluated tools assigned all these bases as unpaired, correctly leaving their WC-edges available for ligand contacts in 50% to 80% of ligand-bound RNA cases in our dataset (Figure 8D). The prediction tools, particularly SPOT-RNA2 and UFold_nc, which demonstrated competitive prediction performance, tend to overpredict base pairing for WC-edge ligand-contacted bases. Overall, WC-edge ligand-contacted bases further complicate the secondary structure prediction within ligand binding sites.

Which Conformation Is Captured by the Prediction, Apo- or Holo-Conformation?

RNA can adopt different base pairing patterns when binding to a ligand. When using secondary structure prediction tools to predict ligand-bound RNAs, it is important to remember that these tools were not designed to specifically predict ligand-bound structures. It is still unclear whether these tools capture the ligand-bound or unbound state of the RNAs. Because corresponding ligand-unbound RNA structures for the ligand-bound RNAs in our dataset are scarce, we focus on four representative examples. These RNAs differ in the levels of conformational change upon ligand binding: (1) the SAM-III

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

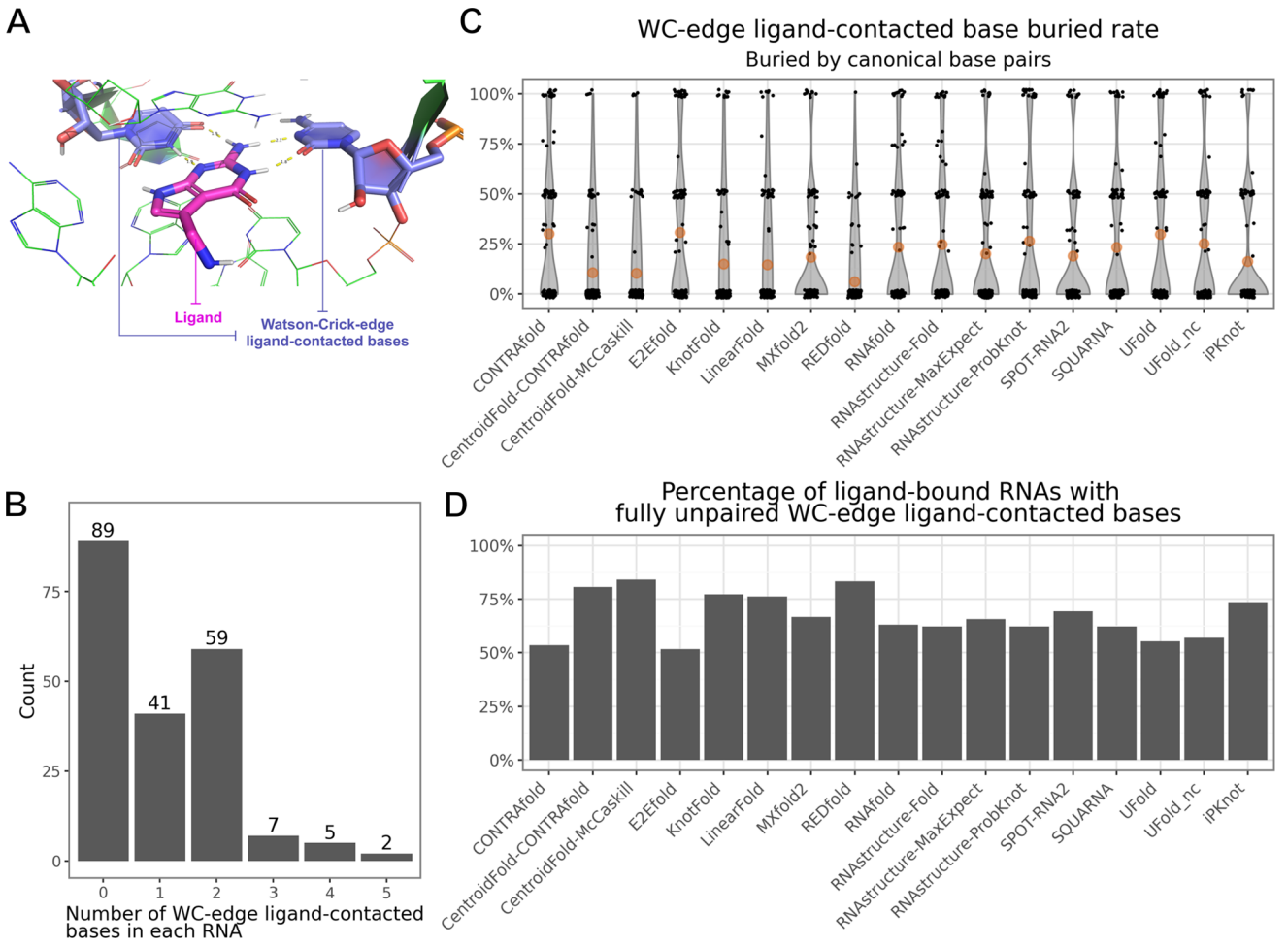


Figure 8. (A) Bases that interact with ligands via hydrogen bonds from their Watson−Crick edges are termed WC-edge ligand-contacted bases throughout this work. (B) Distribution of the number of WC-edge ligand-contacted nucleobases per PDB entry across the dataset. (C) Violin plot showing the buried rate of WC-edge ligand-contacted bases, defined as the percentage of those bases that are incorrectly predicted as paired by each method on the *x*-axis. A value of 0% indicates that all such bases are correctly predicted as unpaired, while 100% indicates that all are incorrectly predicted as paired. The rates for individual ligand-bound RNAs are shown as black dots, and orange circles denote the mean buried rate for each method. (D) Percentage of RNAs for which all WC-edge ligand-contacted bases are predicted as unpaired by each prediction tool.

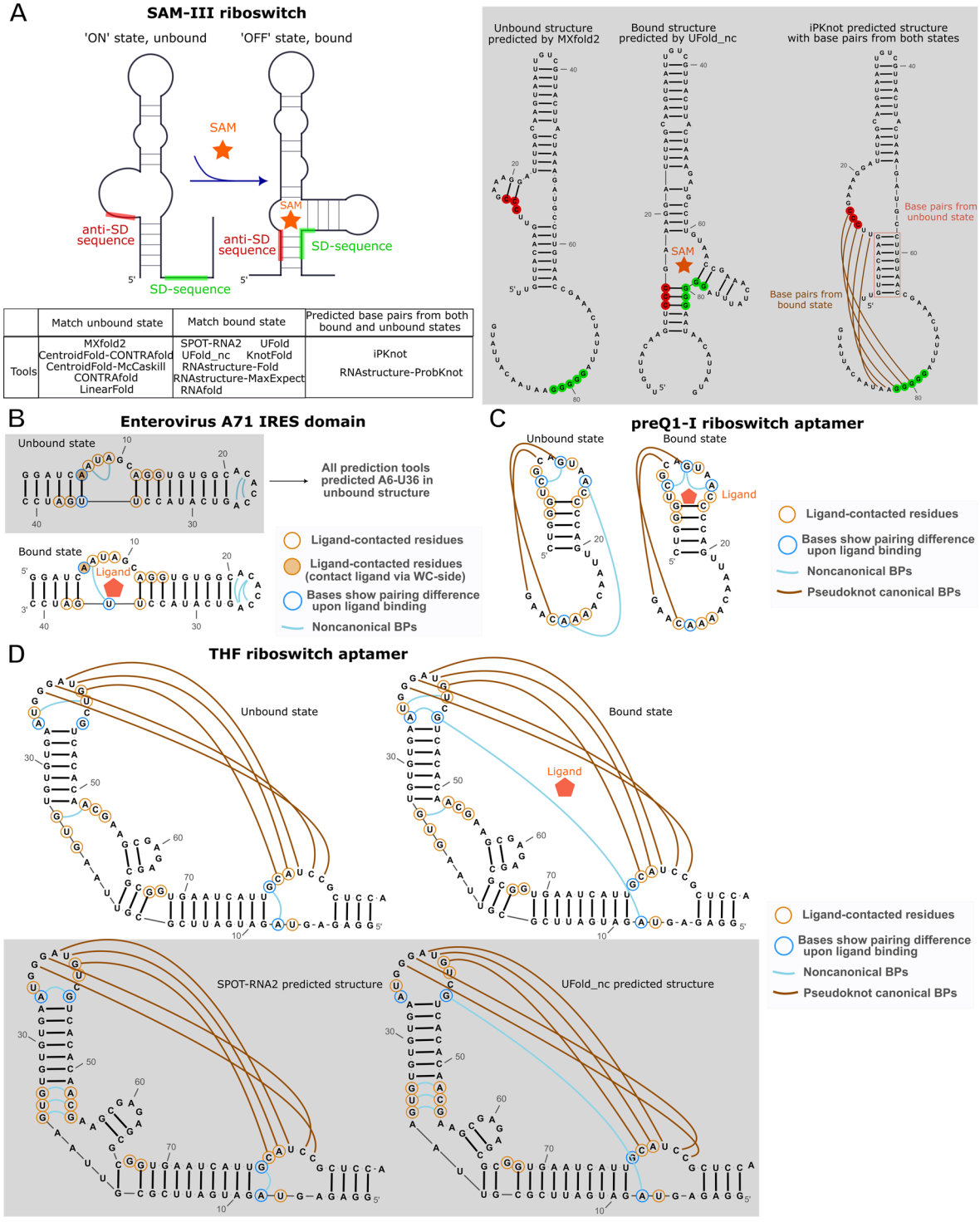
riboswitch undergoes substantial changes in canonical base pairing upon ligand binding; (2) the Internal Ribosome Entry Site (IRES) domain of Enterovirus A71 experiences a single canonical base pair change; and (3) the preQ1-I riboswitch aptamer and the THF riboswitch aptamer display changes only in noncanonical base pairs.

The SAM-III riboswitch recognizes S-adenosyl methionine (SAM) and undergoes a conformational change to regulate the translation of its target gene. In the unbound state, the Shine−Dalgarno (SD) sequence at the 3′ side is unpaired, but it pairs with the upstream anti-SD sequence upon SAM binding (Figure 9A). The ligand-bound structure in PDB entry 3e5c does not include the complete sequence upstream of the anti-SD sequence, so we took the SAM-III riboswitch sequence from*Enterococcus faecalis*, which contains the complete sequence, for our predictions. Seven tools, such as UFold_nc, predicted the bound state of the riboswitch by pairing anti-SD and SD sequences. Five tools, represented by MXfold2, paired the anti-SD sequence with the upstream 5′ side region and exposed the SD sequence, which is more consistent with the unbound state. Tools such as iPKnot predicted intermediate conformations with base pairs from both the bound and unbound states. This indicates the complexity of the possible base pairing patterns in riboswitches. Overall, different conformations of the *Enterococcus faecalis* SAM-III riboswitch were captured by applying multiple secondary structure prediction tools.

The ligand-bound (PDB ID: 6xb7)[82] and unbound (PDB ID: 5v16)[83] structures of the Enterovirus A71 IRES domain differ by a single canonical base pair at the ligand binding site (Figure 9B). Specifically, the A6-U36 base pair is present in the apo structure but is disrupted upon compound binding. All prediction tools, however, assigned the A6-U36 base pair, thereby predicting the secondary structure corresponding to the apo state of the IRES domain.

The ligand-bound preQ1-I and THF riboswitch aptamers (PDB ID: 6vui and 4lvv) have corresponding unbound structures (PDB ID: 6vuh and 7kd1), but these only include the aptamer regions without the expression platform region resolved.[84−86] The unbound structures retain the canonical base pairing pattern of the bound state, and differences are limited to noncanonical base pairs at the binding sites. We evaluated the predicted noncanonical base pairs for these aptamers given by the tools capable of this function. In the preQ1-I riboswitch aptamer, a triplex is present in the unbound structure, and a quartet is present in the bound structure at the

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX



	Match unbound state	Match bound state	Predicted base pairs from both bound and unbound states
Tools	MXfold2 CentroidFold-CONTRAfold CentroidFold-McCaskill CONTRAfold LinearFold	SPOT-RNA2 UFold UFold_nc KnotFold RNAstructure-Fold RNAstructure-MaxExpect RNAfold	iPKnot RNAstructure-ProbKnot

Figure 9. Secondary structure schemes of the discussed RNAs showing conformational changes upon ligand binding. (A) SAM-III riboswitch from *Enterococcus faecalis*, including expression platform with Shine−Dalgarno sequence (green). Base pairing patterns in SAM-unbound and SAM-bound states are shown on the left and the representative predicted structures on the right. (B) Enterovirus A71 IRES domain, where a single canonical base pair (A6-U36) is disrupted upon ligand binding. All prediction tools assigned the A6-U36 pair. (C) PreQ1-I riboswitch aptamer shows noncanonical base pair changes upon ligand binding. (D) THF riboswitch aptamer in ligand-unbound and bound states (top panel), with predicted structures from SPOT-RNA2 and UFold_nc (lower panel). Only noncanonical base pairs present in ligand-contacted regions are annotated.

ligand binding site (Figure 9C). However, none of the tools identified any of these noncanonical base pairs. For the THF riboswitch aptamer, A8 forms a long-distance contact with A34-G44 as a triplex in the bound structure, while in the unbound state, A8 pairs with G78 on the opposite strand. Only UFold_nc and SPOT-RNA2 made predictions involving these noncanonical base pairs: UFold_nc correctly predicted A8-G44 in the bound state, and SPOT-RNA2 identified A8-G78 in the unbound structure as well as A34-G44 in the bound structure (Figure 9D). Further inspection of the 3D structure

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

of the ligand-bound THF riboswitch revealed possible hydrogen bonding between the Hoogsteen edge of A8 and the sugar edge of G78, an interaction not annotated as a noncanonical base pair by FR3D. As a result, SPOT-RNA2 predicted a base pairing pattern for the THF riboswitch aptamer that more closely matches its ligand-bound structure.

DISCUSSION AND CONCLUSIONS

RNA secondary structure prediction provides primary structural insights for novel RNA sequences, which is crucial for RNA-targeted drug discovery. Our work benchmarked widely used RNA secondary structure prediction tools against ligand-bound RNAs. Most evaluated tools achieved a high level of global folding accuracy, with recent methods, UFold, UFold_nc, MXfold2, SPOT-RNA2, and SQUARNA, showing slightly better performance. Interestingly, while machine learning (ML)-based approaches have brought advances to the field,[32−36] those evaluated here did not consistently outperform non-ML-based methods. RNAs with long sequences often adopt multiple base pairing possibilities and complex motifs such as pseudoknots (PK), multiway junctions (MWJ), and G-quadruplexes (GQ).[40,41,44] The accurate prediction of the complex secondary structure motifs from RNA sequence remains challenging for both RNA secondary and 3D structure predictions.[65] Protein binding potentially adds further complexity to secondary structure prediction in these RNAs.[87−89] Consequently, predicting secondary structure—and especially reconstructing ligand binding sites—is challenging for longer RNAs. It is important to note that our benchmarking focused on ligand-bound RNAs with available 3D structures, predominantly short RNAs comprising only 84 clusters based on Rfam families. Therefore, the prediction accuracy observed in our study is generally good and may differ from results reported in previous benchmarking studies using larger and more diverse datasets.

PKs are common RNA motifs that determine the global folding and can contribute to ligand binding site formation.[29,42,43] However, even the best-performing prediction tools struggle to accurately identify PKs. This limitation is partly due to insufficient thermodynamic data for PKs and the algorithmic challenges inherent to their prediction. Accurate recognition of PKs depends on correct prediction of the global base pairing pattern, since it must capture non-nested base pairing rather than just the pairs within PK regions. Collectively, current RNA secondary structure prediction tools are generally unable to reconstruct ligand binding sites in PK-containing RNAs.

Although secondary structure prediction tools can sometimes successfully reconstruct RNA ligand binding sites, it is generally impractical to rely on a single tool for this purpose. Prediction accuracy at ligand binding sites is lower than that for the entire RNA length. More critically, the tools rarely capture the local complex secondary structure elements that are key to shaping the binding pockets. Ligand-contacted nucleotides in double helices and hairpin loops are identified with reasonable accuracy. However, internal loops frequently contribute to the formation of ligand binding sites; yet, they are rarely predicted correctly. Similarly, poor recognition of PKs results in inaccurate prediction of pseudoknot loops, which commonly dominate ligand binding sites in PK-containing RNAs.

Noncanonical base pairs are important for RNA folding and shaping ligand binding sites and frequently occur in ligand-contacted regions in various forms, i.e., single noncanonical base pairs, triplexes, G-quartets, and other multiplexes.[49,50] Traditional prediction tools do not account for noncanonical base pairs, and those developed to identify them demonstrate low overall prediction accuracy; recognition of associated structural elements within ligand-contacted regions is even less reliable. Several specialized prediction tools, such as BayesPairing 2 and RNAMoIP, have also been developed to learn and identify RNA modules with noncanonical base pairs that share local sequence similarity.[90−92] However, the accuracy for noncanonical base pair prediction remains much lower than for canonical base pairs. Current secondary structure prediction methods are far from providing accurate details of noncanonical base pairing at ligand binding sites.[61,93] This is largely due to the scarcity of annotated noncanonical base pairs—such knowledge requires resolved RNA 3D structures, which remain limited.[94−96] Consequently, it is currently impractical to train computational models that can reliably predict noncanonical interactions. Although 3D structure prediction from sequence and secondary structure can sometimes recover noncanonical base pairs, resolved RNA structures are needed to enrich noncanonical base pair annotation.[97,98]

Lastly, it is critical to assign Watson−Crick (WC) edge ligand-contacted bases as unpaired to expose their WC edge for ligand interactions. Although prediction tools rarely block the WC edges of these bases, any such errors can further hinder the accurate reconstruction of RNA ligand binding sites.

The inherent flexibility of RNA permits alternative base pairing possibilities within its sequence. Upon ligand binding, RNAs may remain mostly static or undergo conformational changes ranging from subtle shifts in noncanonical base pairs to major rearrangements of canonical base pairing. In particular, riboswitches and RNA aptamers vastly change their base pairings in response to ligand binding.[62−64] Unfortunately, RNA secondary structure prediction tools were designed to predict the RNA structure without any concern of ligand binding. As illustrated by the predictions of the full-length SAM-III riboswitch sequence—including both its aptamer and expression platform regions—it is possible to apply different prediction tools to sample both ligand-bound and unbound conformations. Utilizing multiple secondary structure prediction tools can be a strategy to explore alternative base pairing patterns contributing to RNA dynamics. However, selecting the biologically relevant conformations from the predictions is a follow-up challenge. Additional computational and experimental validations are needed to determine whether any of the predicted models resemble the ligand-bound state. According to our limited examples of ligand-bound and unbound RNA, we find that prediction tools fail to capture small yet functionally relevant changes in both canonical and noncanonical base pairs induced by ligand binding. This underscores the risks of relying solely on secondary structure prediction methods for identifying druggable sites in RNA, which may compromise downstream RNA-targeted drug discovery efforts.

Our study provides a comprehensive overview of the strengths and limitations of incorporating secondary structure prediction tools into RNA-targeted drug discovery pipelines. While certain evaluated tools showed advances for some ligand-bound RNAs, a universally reliable predictor does not exist for predicting the global folding of all of them. Beyond classical and ML-based prediction methods, integrating

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. RNA Secondary Prediction Tools Benchmarked for Ligand-Bound RNAs

RNA secondary structure prediction tools	Machine learning-based	Pseudoknot prediction	Noncanonical base pairing prediction	Note for usage
RNAfold[26,27]	No	No	No	-
RNAstructure-Fold[28,29]	No	No	No	-
RNAstructure-MaxExpect[28,66]	No	No	No	-
RNAstructure-ProbKnot[28,44]	No	Yes	No	-
LinearFold[67]	No	No	No	-
IPKnot[45,46]	No	Yes	No	-
SPOT-RNA2[a52,53]	Yes	Yes	Yes	32 GB RAM, 500 GB disk space are suggested to support the in-memory operations for RNA sequence length less than 500. Multiple CPU threads are also recommended.
CentroidFold-CONTRAfold[68,69]	No	No	Yes	-
CentroidFold-McCaskill[68,70]	No	No	Yes	-
CONTRAfold[69]	Yes	No	No	-
KnotFold[51]	Yes	Yes	Yes	-
REDfold[71]	Yes	Yes	Yes	Input sequence length up to 720 nucleotides. Do not accept input sequence with "N".
E2Efold[72]	Yes	Yes	Yes	-
UFold[33]	Yes	Yes	No	Input sequence length up to 500 nucleotides.
UFold_nc[b33]	Yes	Yes	Yes	Input sequence length up to 500 nucleotides.
MXfold2[32,36]	Yes	No	No	-
SQUARNA[47]	No	Yes	No	-

[a]SPOT-RNA2 is the only multiple-sequence alignment-based method evaluated here. [b]UFold_nc refers to UFold tool with noncanonical base pair prediction activated ("−nc True" option).

evolutionary information and experimental data into the secondary structure is recommended to improve prediction accuracy for these RNAs.[59,60] Recently, language model (LM)-based RNA secondary structure prediction tools have shown competitiveness with classical and ML-based approaches. Some can capture homologous information from single sequence input, avoiding the time-consuming multiple sequence alignment step. Nevertheless, the accuracy of both ML-based and LM-based approaches generally depends the homology between the predicted sequence and their training datasets.[32,38,99] Identification and reconstruction of ligand binding sites from novel RNA sequences need high prediction accuracy. Currently, secondary structure prediction tools show inadequate performance in ligand binding site reconstruction. Ideally, specialized prediction tools that can identify pseudoknots, noncanonical base pair modules, or locate ligand binding sites based on either sequence or secondary structures can be integrated to improve the structural detail of RNA models.[100,101] However, RNA structures are stabilized not only by maximizing base pairings but also largely by a 3D context, such as tertiary contacts. Therefore, downstream computational strategies that leverage 3D information, such as RNA 3D structure prediction, virtual screening, and molecular dynamics simulations, are essential to evaluate and refine the secondary structure prediction outputs. Such techniques may improve the reconstruction of both canonical and noncanonical base pairing patterns within ligand binding sites.[98,102−104] Besides, continued efforts to resolve new RNA 3D structures are critical for expanding RNA structural dataset and deepening our understanding of RNA folding principles. Ultimately, this will support the development of new prediction tools capable of identifying a more complete network of canonical and noncanonical base pairs. In parallel, ongoing advances in artificial intelligence, including both cutting-edge deep learning-based and LM-based methods, may further strengthen the role of RNA secondary structure prediction in the rapid screening of druggable RNA sites.

MATERIALS AND METHODS

Ligand-Bound RNA Structure Dataset Curation

This study benchmarked RNA secondary structure prediction tools against ligand-bound RNAs and aimed to investigate the predictions on the ligand binding sites. For these purposes, we collected PDBs of ligand-bound RNAs from HARIBOSS,[23] the dataset curated for ligand-bound RNA 3D structure prediction benchmarking work from Nithin et al.,[65] and newly submitted PDBs in EMBL PDBe until 4 September 2024.[22,105] The ligand-RNA complexes that fall into the following categories were discarded: (1) the RNA molecule has more than 1 chain or has a broken chain where multiple nucleotides are missing; (2) sequence length is over 1000 nucleotides; (3) crystallization buffer compounds are the only ligands in the complex; (4) RNA functions as a multimer. The filtering criteria were informed by the need to benchmark against ligand-bound RNAs with biologically functional ligands and meet the requirements of RNA secondary structure prediction tools. Generally, RNA secondary structure prediction cannot support the accurate prediction of long sequences (typically those exceeding 1000 nucleotides) and multi-chain RNAs.[24,106,107] Note that some crystal structures contain homomultimer RNA chains in the asymmetric unit, even though the RNA is functionally annotated as a monomer. In this case, we did not discard the entry but instead retained the first RNA chain with the ligand in its binding site. For the complexes resolved by NMR, we included only the first model from each PDB entry in the dataset. The filtered dataset was subsequently clustered by RNA family using Rfam covariance models and the Infernal program, with each cluster corresponding to an Rfam family.[108,109] For RNAs without an assigned Rfam family, e.g., RNA aptamers, we applied the CD-HIT-EST program with an 80% sequence identity threshold to complete the clustering.[110,111] The sequence identity threshold was selected to match the value used for dataset division during the training of the evaluated machine learning-based tools. Within each cluster,

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

complexes containing identical ligands in the same binding site were considered redundant, and only the structure with the highest resolution was retained. The dataset construction protocol yielded 84 clusters comprising 203 ligand-RNA complexes (Table S1). Of the 203 complexes, 164 originated from HARIBOSS, 21 from Nithin's dataset, and 18 from newly deposited PDB entries. It is noteworthy that 106 of the 203 ligand-bound RNAs are riboswitches and 62 are aptamers, indicating a functional bias within the dataset (Table S1 and Figure S7). All PDB files in the dataset went through a cleaning process using pdb-tools[112] and in-house scripts in order to remove ions, bound proteins, functionally irrelevant ligands, renumber RNA nucleotides, etc. We extracted the sequence and secondary structure information on each complex using FR3D as ground-truth annotations, a program recognized for its accuracy in base pair annotation.[113] We also annotated secondary structure using the x3dna-dssr program v2.4.2 to independently validate our results.[114,115] While these annotation methods might yield different base pairing patterns, consistent conclusions drawn from both methods strengthen the reliability of our analysis. The modified nucleobases were relabeled as their corresponding canonical bases in the sequence. The PDB 6e1v and 7e9e have several unpaired nucleobases unresolved in the cocrystallized structures, which were denoted as "N" in sequence and unpaired in secondary structure. In addition to canonical base pairs, we also obtained noncanonical base pairs from both FR3D and x3dna-dssr outputs.

We further enriched the information in the dataset using the PDBe API developed by the European Molecular Biology Laboratory of the European Bioinformatics Institute (EMBL-EBI),[105] including sequence length, protein-binding information, etc. The RNA motif annotations, e.g., multiway junctions (MWJ), pseudoknots (PK), and G-quadruplexes (GQ), were obtained from the extracted RNA secondary structures. For each RNA in the dataset, we identified and annotated secondary structure elements and noncanonical base pair-associated structural modules—hereafter collectively referred to as structure elements—for all bases, which are based on and expanded from Turner's nearest-neighbor model.[30,31] As illustrated in Figure 1B, these structure elements include those only derived from canonical base pairing, such as helices, bulge loops, exterior loops, hairpin loops, internal loops, multibranch loops, and multiway junction BP ("BP" refers to "base pair"), as well as elements associated with noncanonical base pairs: G-quartet, G-quartet loop (GQ loop), single noncanonical base pair, triplex, and multiplex. These annotations were performed using an in-house script. Importantly, we additionally defined two loop elements not covered by Turner's model. Pseudoknot loops are loops flanked on both sides by non-nested base pairs that define a pseudoknot, whereas linkage loops are loops that do not belong to pseudoknot loops and cannot be enclosed by any base pairs or any other loops. We further extracted ligand binding site information by in-house scripts using the Biopython package:[116] (1) ligand-contacted nucleotides for each complex, defined by a distance threshold of 6 Å to any ligand atoms; (2) the nucleotides that show H-bonding interactions with ligands via their Watson−Crick (WC) nucleobase edges (as WC-edge ligand-contacted bases, see Figure 8A). The dataset with the annotated details is summarized in Table S1, and we also provided structural information including sequence, secondary structures, noncanonical base pairs, and WC-edge ligand-contacted bases in Zenodo (DOI: 10.5281/zenodo.18221787).

Benchmarking RNA Secondary Structure Prediction Tools

Seventeen commonly used prediction tools were included in this benchmarking study and are summarized in Table 1. The benchmarked prediction tools employ a range of methodologies, from classical approaches combining Turner's nearest-neighbor model with dynamic programming (e.g., RNAfold) to advanced ML-based methods (e.g., UFold). All tools are based on single sequence analysis, except SPOT-RNA2, which incorporates multiple sequence alignment-based analyses and demonstrates improved performance on sequences with available evolutionary information compared to the single sequence-based SPOT-RNA.[52,53] The tools also show different capabilities in PK and noncanonical base pairing predictions. We ran RNA secondary structure prediction for all tools on our dataset locally and used their default setups. For more details about the sequence input requirements, see Table 1.

Notably, the machine learning-based prediction tools were trained on specific datasets, which may lead to overfitting and reduced accuracy for RNA families not presented in their training sets.[32,99] To assess the risk of overfitting for the evaluated tools on our dataset, we independently compared sequence identities between our dataset and the training set of each machine learning-based tool. Sequence identity assessment was performed using CD-HIT-EST-2D, with an 80% threshold to define overlap.[110,111] The sequence identity threshold has the same value as that used for dataset clustering as described above. For each RNA in our dataset, we annotated whether it overlaps with the training data of any of these tools. We acknowledge that various methods are available to determine whether RNAs in our dataset are included in the training sets of machine learning-based tools, such as using the Infernal protocol applied in our dataset clustering. However, for rapid evaluation, we chose to use the sequence identity. We used RNAcanvas for visual comparison of reference and predicted secondary structures.[117]

Performance Evaluation Metrics

To evaluate the accuracy of predicted secondary structures, we used the F1-score as suggested by Mathews,[118] which balances precision (PPV) and sensitivity (SN) via the harmonic mean (eq 1), and we applied the Accuracy (ACC) metric as defined in the work of REDfold (eq 2).[71] The definitions of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are provided in Table 2. Note that the F1-score and ACC metrics do not take noncanonical base pairs into account. Noncanonical base pairs present in the reference structures, as well as those predicted in noncanonical combinations by the applicable tools, were excluded from F1-score and ACC value calculations. This assessment approach prevents predicted noncanonical base pairs from being counted as false negatives, which could otherwise deteriorate the evaluations of canonical base pair predictions.

Table 2. Definitions of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) in This RNA Secondary Structure Prediction Benchmarking Study

TP	Number of bases correctly predicted as paired **and** correctly paired with another base.
FP	Number of bases predicted as paired, but they are actually unpaired **or** paired with a wrong base.
FN	Number of bases predicted as unpaired, but they are actually paired.
TN	Number of bases correctly predicted as unpaired.

$$F1 = \frac{2PPV \cdot SN}{PPV + SN} = \frac{2TP}{2TP + FN + FP} \quad (1)$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

F1-score and ACC values were calculated for each ligand-bound RNA predicted by each tool. To mitigate bias from varying cluster sizes, these metrics were subsequently averaged within each cluster for the evaluations.

The performance of PK recognition was judged by the strict PK base pair hit rate. Specifically, we converted the prediction output into dot-bracket format without the consideration of predicted noncanonical base pairs, from which we identified the base pairs within PK regions and PK orders from bracket types ("[]", "{}", or "< >"). The numbers of correctly predicted base pairs within PK regions under different PK orders were counted and subsequently divided by the total number of base pairs for each PK order, yielding the strict PK base pair hit rate.

To evaluate the prediction accuracy specifically within ligand-contacted regions, we calculated ACC values for these regions,

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

reflecting the prediction hit rate for ligand-contacted nucleotides. Because of the complex local structures within ligand-contacted regions, it is also important for prediction tools to reconstruct relevant structural elements rather than merely identifying base-pairing patterns. Accordingly, we grouped ligand-contacted nucleotides by their structure elements and calculated the hit rate for each structure element within these regions. Additionally, since predicted secondary structures may assign base pairing to the WC-edge ligand-contacted bases—potentially burying the interaction surfaces—we estimated the buried rate by calculating the probability that these bases were predicted as paired.

For the evaluation of prediction accuracy on noncanonical base pairs, we counted all predicted base pairs that overlapped with noncanonical base pairs in the reference structures as hits. The noncanonical base pair hit rate was then calculated by dividing the number of hits by the total number of noncanonical base pairs in the reference structures.

■ ASSOCIATED CONTENT

Data Availability Statement

All structural information in our curated dataset is available in Zenodo (DOI: 10.5281/zenodo.18221787) and in supplementary data online.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.6c00108.

Ligand-bound RNA dataset details, overview of secondary structure annotations, and supplementary benchmarking results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Zhengyue Zhang − *Medicinal Chemistry R&I, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg 431 83, Sweden;* orcid.org/0000-0002-5322-7087; Email: zhengyue.zhang@astrazeneca.com

Authors

Gaia Dolcetti − *Medicinal Chemistry R&I, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg 431 83, Sweden; Department of Biomedicine, University of Bergen, Bergen 5020, Norway;* orcid.org/0009-0005-3826-2676

Christian Tyrchan − *Medicinal Chemistry R&I, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg 431 83, Sweden;* orcid.org/0000-0002-6470-984X

Leonardo De Maria − *Medicinal Chemistry R&I, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg 431 83, Sweden;* orcid.org/0000-0002-8061-4242

Giovanni Bussi − *Scuola Internazionale Superiore di Studi Avanzati, Trieste 34136, Italy;* orcid.org/0000-0001-9216-5782

Werngard Czechtizky − *Medicinal Chemistry R&I, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg 431 83, Sweden*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.6c00108

Author Contributions

Z.Z. curated the dataset, installed and ran programs, analyzed results, and wrote the manuscript; G.D. analyzed results and wrote the manuscript; C.T. and L.D.M. supervised the analysis and wrote the manuscript; W.C. and G.B. supervised the study. All authors have given approval to the final version of the manuscript.

Funding

This work has been supported by AstraZeneca and received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement 101168667.

Notes

The authors declare the following competing financial interest(s): Zhengyue Zhang is enrolled as a postdoctoral fellow in AstraZeneca AB R&D, and Gaia Dolcetti is pursuing an industrial PhD held by AstraZeneca AB R&D and University of Bergen in Norway. Christian Tyrchan, Leonardo De Maria and Werngard Czechtizky are employees of AstraZeneca AB R&D and hold shares of the company. The authors report no other conflicts of interest in this work.

■ ACKNOWLEDGMENTS

We thank Dr. Thomas Löhr for supporting MAIZE (https://molecularai.github.io/maize/), a graph-based workflow manager, during our benchmarking work.

■ ABBREVIATIONS

ML, machine learning; LM, language model; PK, pseudoknot; MWJ, multiway junction; GQ, G-quartet; WC, Watson−Crick

■ REFERENCES

(1) Djebali, S.; Davis, C. A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription in human cells. *Nature* **2012**, *489* (7414), 101−108.

(2) Sharp, P. A. The Centrality of RNA. *Cell* **2009**, *136* (4), 577−580.

(3) Mattick, J. S.; Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **2006**, *15*, R17−R29.

(4) Tsai, M.-C.; Manor, O.; Wan, Y.; Mosammaparast, N.; Wang, J. K.; Lan, F.; Shi, Y.; Segal, E.; Chang, H. Y. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* **2010**, *329* (5992), 689−693.

(5) He, L.; Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **2004**, *5* (7), 522−531.

(6) Neueder, A. RNA-Mediated Disease Mechanisms in Neurodegenerative Disorders. *J. Mol. Biol.* **2019**, *431* (9), 1780−1791.

(7) Jonkhout, N.; Tran, J.; Smith, M. A.; Schonrock, N.; Mattick, J. S.; Novoa, E. M. The RNA modification landscape in human disease. *RNA* **2017**, *23* (12), 1754−1769.

(8) Chen, S.; Mao, Q.; Cheng, H.; Tai, W. RNA-Binding Small Molecules in Drug Discovery and Delivery: An Overview from Fundamentals. *J. Med. Chem.* **2024**, *67* (18), 16002−16017.

(9) Lin, J.; Zhou, D.; Steitz, T. A.; Polikanov, Y. S.; Gagnon, M. G. Ribosome-Targeting Antibiotics: Modes of Action, Mechanisms of Resistance, and Implications for Drug Design. *Annu. Rev. Biochem.* **2018**, *87*, 451−478.

(10) Singh, R. N.; Ottesen, E. W.; Singh, N. N. The First Orally Deliverable Small Molecule for the Treatment of Spinal Muscular Atrophy. *Neurosci Insights* **2020**, *15*, 2633105520973985.

(11) Lee, E. R.; Blount, K. F.; Breaker, R. R. Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression. *RNA Biol.* **2009**, *6* (2), 187−194.

(12) Ratni, H.; Ebeling, M.; Baird, J.; Bendels, S.; Bylund, J.; Chen, S. K.; Denk, N.; Feng, Z.; Green, L.; Guerard, M.; et al. Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 *SMN2* Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA). *J. Med. Chem.* **2018**, *61* (15), 6501−6517.

(13) Naryshkin, N. A.; Weetall, M.; Dakka, A.; Narasimhan, J.; Zhao, X.; Feng, Z.; Ling, K. K. Y.; Karp, G. M.; Qi, H.; Woll, M. G.; et al. *SMN2* splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. *Science* **2014**, *345* (6197), 688−693.
(14) Crooke, S. T.; Witztum, J. L.; Bennett, C. F.; Baker, B. F. RNA-Targeted Therapeutics. *Cell Metab.* **2018**, *27* (4), 714−739.
(15) Childs-Disney, J. L.; Yang, X.; Gibaut, Q. M. R.; Tong, Y.; Batey, R. T.; Disney, M. D. Targeting RNA structures with small molecules. *Nat. Rev. Drug Discovery* **2022**, *21* (10), 736−762.
(16) Kaur, J.; Sharma, A.; Mundlia, P.; Sood, V.; Pandey, A.; Singh, G.; Barnwal, R. P. RNA−Small-Molecule Interaction: Challenging the "Undruggable" Tag. *J. Med. Chem.* **2024**, *67* (6), 4259−4297.
(17) Taghavi, A.; Springer, N. A.; Zanon, P. R. A.; Li, Y.; Li, C.; Childs-Disney, J. L.; Disney, M. D. The evolution and application of RNA-focused small molecule libraries. *RSC Chem. Biol.* **2025**, *6* (4), 510−527.
(18) Warner, K. D.; Hajdin, C. E.; Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discovery* **2018**, *17* (8), 547−558.
(19) Tessaro, F.; Scapozza, L. How 'Protein-Docking' Translates into the New Emerging Field of Docking Small Molecules to Nucleic Acids? *Molecules* **2020**, *25* (12), 2749.
(20) Matthes, F.; Massari, S.; Bochicchio, A.; Schorpp, K.; Schilling, J.; Weber, S.; Offermann, N.; Desantis, J.; Wanker, E.; Carloni, P.; et al. Reducing Mutant Huntingtin Protein Expression in Living Cells by a Newly Identified RNA CAG Binder. *ACS Chem. Neurosci.* **2018**, *9* (6), 1399−1408.
(21) Big pharma craves slice of AI-based RNA drug discovery. *Nat. Biotechnol.* **2023**, *41* (3), 305−305..
(22) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. D.; Christie, C.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47* (D1), D520−D528.
(23) Panei, F. P.; Torchet, R.; Ménager, H.; Gkeka, P.; Bonomi, M. HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design. *Bioinformatics* **2022**, *38* (17), 4185−4193.
(24) Mathews, D. H.; Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **2006**, *16* (3), 270−278.
(25) Wang, D.; Jiang, Y.; He, L.; Zhang, L.; Zhou, R.; Zhang, D.Cross-talk between RNA secondary and three-dimensional structure prediction: a comprehensive study*bioRxiv***2025**
(26) Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol Biol.* **2011**, *6* (1), 26.
(27) Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **1994**, *125* (2), 167−188.
(28) Reuter, J. S.; Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf.* **2010**, *11* (1), 129.
(29) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **1999**, *288* (5), 911−940.
(30) Schroeder, S. J.; Turner, D. H. Optical melting measurements of nucleic acid thermodynamics. *Methods Enzymol.* **2009**, *468*, 371−387.
(31) Turner, D. H.; Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D280−D282.
(32) Sato, K.; Akiyama, M.; Sakakibara, Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **2021**, *12* (1), 941.
(33) Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; Xie, X. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **2022**, *50* (3), No. e14.
(34) Sato, K.; Hamada, M. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Briefings Bioinf.* **2023**, *24* (4), 1−13.
(35) Chaturvedi, M.; Rashid, M. A.; Paliwal, K. K. RNA structure prediction using deep learning — A comprehensive review. *Comput. Biol. Med.* **2025**, *188*, 109845.
(36) Akiyama, M.; Sato, K.; Sakakibara, Y. A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *J. Bioinform Comput. Biol.* **2018**, *16* (6), 1840025.
(37) Giuseppe Sacco, G. B.; Sanguinetti, G.Machine Learning for RNA Secondary Structure Prediction: a review of current methods and challenges*arXiv***2025**
(38) Zablocki, L. I.; Bugnon, L. A.; Gerard, M.; Di Persia, L.; Stegmayer, G.; Milone, D. H. Comprehensive benchmarking of large language models for RNA secondary structure prediction. *Brief Bioinform.* **2025**, *26* (2), bbaf137.
(39) Penic, R. J.; Vlasic, T.; Huber, R. G.; Wan, Y.; Sikic, M. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nat. Commun.* **2025**, *16* (1), 5671.
(40) Lyngsø, R. B.; Pedersen, C. N. S. RNA Pseudoknot Prediction in Energy-Based Models. *J. Comput. Biol.* **2000**, *7* (3−4), 409−427.
(41) Akutsu, T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.* **2000**, *104* (1−3), 45−62.
(42) Leonard, C. W.; Hajdin, C. E.; Karabiber, F.; Mathews, D. H.; Favorov, O.; Dokholyan, N. V.; Weeks, K. M. Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry* **2013**, *52* (4), 588−595.
(43) Staple, D. W.; Butcher, S. E. Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biol.* **2005**, *3* (6), No. e213.
(44) Bellaousov, S.; Mathews, D. H. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA* **2010**, *16* (10), 1870−1880.
(45) Sato, K.; Kato, Y.; Hamada, M.; Akutsu, T.; Asai, K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **2011**, *27* (13), i85−i93.
(46) Poolsap, U.; Kato, Y.; Akutsu, T. Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC Bioinf.* **2009**, *10* (S1), S38.
(47) Bohdan, D. R.; Nikolaev, G. I.; Bujnicki, J. M.; Baulin, E. F.SQUARNA - an RNA secondary structure prediction method based on a greedy stem formation model*bioRxiv***2024**
(48) Gray, M.; Trinity, L.; Stege, U.; Ponty, Y.; Will, S.; Jabbari, H. CParty: hierarchically constrained partition function of RNA pseudoknots. *Bioinformatics* **2024**, *41* (1), btae748.
(49) Oliver, C.; Mallet, V.; Gendron, R. S.; Reinharz, V.; Hamilton, W. L.; Moitessier, N.; Waldispuhl, J. Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic Acids Res.* **2020**, *48* (14), 7690−7699.
(50) Kligun, E.; Mandel-Gutfreund, Y. Conformational readout of RNA by small ligands. *RNA Biol.* **2013**, *10* (6), 981−989.
(51) Gong, T.; Ju, F.; Bu, D. Accurate prediction of RNA secondary structure including pseudoknots through solving minimum-cost flow with learned potentials. *Commun. Biol.* **2024**, *7* (1), 297.
(52) Singh, J.; Hanson, J.; Paliwal, K.; Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **2019**, *10* (1), 5407.
(53) Singh, J.; Paliwal, K.; Zhang, T.; Singh, J.; Litfin, T.; Zhou, Y. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics* **2021**, *37* (17), 2589−2600.
(54) Tagashira, M.; Asai, K. ConsAlifold: considering RNA structural alignments improves prediction accuracy of RNA consensus secondary structures. *Bioinformatics* **2022**, *38* (3), 710−719.

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(55) Wu, Y.; Shi, B.; Ding, X.; Liu, T.; Hu, X.; Yip, K. Y.; Yang, Z. R.; Mathews, D. H.; Lu, Z. J. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.* **2015**, *43* (15), 7247−7259.
(56) Spasic, A.; Assmann, S. M.; Bevilacqua, P. C.; Mathews, D. H. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res.* **2018**, *46* (1), 314−323.
(57) Siegfried, N. A.; Busan, S.; Rice, G. M.; Nelson, J. A. E.; Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **2014**, *11* (9), 959−965.
(58) Mathews, D. H.; Disney, M. D.; Childs, J. L.; Schroeder, S. J.; Zuker, M.; Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (19), 7287−7292.
(59) Wirecki, T. K.; Merdas, K.; Bernat, A.; Boniecki, M. J.; Bujnicki, J. M.; Stefaniak, F. RNAProbe: a web server for normalization and analysis of RNA structure probing data. *Nucleic Acids Res.* **2020**, *48* (W1), W292−W299.
(60) Gumna, J.; Zok, T.; Figurski, K.; Pachulska-Wieczorek, K.; Szachniuk, M. RNAthor - fast, accurate normalization, visualization and statistical analysis of RNA probing data resolved by capillary electrophoresis. *PLoS One* **2020**, *15* (10), No. e0239287.
(61) Justyna, M.; Antczak, M.; Szachniuk, M. Machine learning for RNA 2D structure prediction benchmarked on experimental data. *Briefings Bioinf.* **2023**, *24* (3), 1−9.
(62) Umuhire Juru, A.; Patwardhan, N. N.; Hargrove, A. E. Understanding the Contributions of Conformational Changes, Thermodynamics, and Kinetics of RNA−Small Molecule Interactions. *ACS Chem. Biol.* **2019**, *14* (5), 824−838.
(63) Haller, A.; Soulière, M. F.; Micura, R. The Dynamic Nature of RNA as Key to Understanding Riboswitch Mechanisms. *Acc. Chem. Res.* **2011**, *44*, 1339−1348.
(64) Ha, T.; Zhuang, X.; Kim, H. D.; Orr, J. W.; Williamson, J. R.; Chu, S. Ligand-induced conformational changes observed in single RNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (16), 9077−9082.
(65) Nithin, C.; Kmiecik, S.; Błaszczyk, R.; Nowicka, J.; Tuszyńska, I. Comparative analysis of RNA 3D structure prediction methods: towards enhanced modeling of RNA−ligand interactions. *Nucleic Acids Res.* **2024**, *52* (13), 7465−7486.
(66) Lu, Z. J.; Gloor, J. W.; Mathews, D. H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **2009**, *15* (10), 1805−1813.
(67) Huang, L.; Zhang, H.; Deng, D.; Zhao, K.; Liu, K.; Hendrix, D. A.; Mathews, D. H. LinearFold: linear-time approximate RNA folding by 5′-to-3′ dynamic programming and beam search. *Bioinformatics* **2019**, *35* (14), i295−i304.
(68) Hamada, M.; Kiryu, H.; Sato, K.; Mituyama, T.; Asai, K. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* **2009**, *25* (4), 465−473.
(69) Do, C. B.; Woods, D. A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22* (14), No. e90−e98.
(70) Mccaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **1990**, *29* (6−7), 1105−1119.
(71) Chen, C.-C.; Chan, Y.-M. REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. *BMC Bioinf.* **2023**, *24* (1), 122.
(72) Chen, X.; Yu, L.; Ramzan, U.; Xin, G.; Le, S.RNA Secondary Structure Prediction By Learning Unrolled Algorithms*arXiv***2020**
(73) Leontis, N. B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7* (4), 499−512.
(74) Nagaswamy, U. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.* **2000**, *28* (1), 375−376.
(75) Halder, S.; Bhattacharyya, D. RNA structure and dynamics: A base pairing perspective. *Prog. Biophys. Mol. Biol.* **2013**, *113* (2), 264−283.
(76) Lescoute, A.; Westhof, E. The interaction networks of structured RNAs. *Nucleic Acids Res.* **2006**, *34* (22), 6587−6604.
(77) Butcher, S. E.; Pyle, A. M. The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Acc. Chem. Res.* **2011**, *44* (12), 1302−1311.
(78) Serganov, A.; Patel, J. D. Metabolite Recognition Principles and Molecular Mechanisms Underlying Riboswitch Function. *Annu. Rev. Biophys.* **2012**, *41* (1), 343−370.
(79) David-Eden, H.; Mankin, S. A.; Mandel-Gutfreund, Y. Structural signatures of antibiotic binding sites on the ribosome. *Nucleic Acids Res.* **2010**, *38* (18), 5982−5994.
(80) Padroni, G.; Patwardhan, N. N.; Schapira, M.; Hargrove, A. E. Systematic analysis of the interactions driving small molecule−RNA recognition. *RSC Med. Chem.* **2020**, *11* (7), 802−813.
(81) Seelam, P. P.; Mitra, A.; Sharma, P. Pairing interactions between nucleobases and ligands in aptamer: ligand complexes of riboswitches: crystal structure analysis, classification, optimal structures, and accurate interaction energies. *RNA* **2019**, *25* (10), 1274−1290.
(82) Davila-Calderon, J.; Patwardhan, N. N.; Chiu, L. Y.; Sugarman, A.; Cai, Z.; Penutmutchu, S. R.; Li, M. L.; Brewer, G.; Hargrove, A. E.; Tolbert, B. S. IRES-targeting small molecule inhibits enterovirus 71 replication via allosteric stabilization of a ternary complex. *Nat. Commun.* **2020**, *11* (1), 4775.
(83) Tolbert, M.; Morgan, C. E.; Pollum, M.; Crespo-Hernandez, C. E.; Li, M. L.; Brewer, G.; Tolbert, B. S. HnRNP A1 Alters the Structure of a Conserved Enterovirus IRES Domain to Stimulate Viral Translation. *J. Mol. Biol.* **2017**, *429* (19), 2841−2858.
(84) Schroeder, G. M.; Dutta, D.; Cavender, C. E.; Jenkins, J. L.; Pritchett, E. M.; Baker, C. D.; Ashton, J. M.; Mathews, D. H.; Wedekind, J. E. Analysis of a preQ1-I riboswitch in effector-free and bound states reveals a metabolite-programmed nucleobase-stacking spine that controls gene regulation. *Nucleic Acids Res.* **2020**, *48* (14), 8146−8164.
(85) Trausch, J. J.; Batey, R. T. A disconnect between high-affinity binding and efficient regulation by antifolates and purines in the tetrahydrofolate riboswitch. *Chem. Biol.* **2014**, *21* (2), 205−216.
(86) Wilt, H. M.; Yu, P.; Tan, K.; Wang, Y. X.; Stagno, J. R. Tying the knot in the tetrahydrofolate (THF) riboswitch: A molecular basis for gene regulation. *J. Struct Biol.* **2021**, *213* (1), 107703.
(87) Mohr, S.; Stryker, J. M.; Lambowitz, A. M. A DEAD-Box Protein Functions as an ATP-Dependent RNA Chaperone in Group I Intron Splicing. *Cell* **2002**, *109* (6), 769−779.
(88) Caprara, M. G.; Lehnert, V.; Lambowitz, A. M.; Westhof, E. A Tyrosyl-tRNA Synthetase Recognizes a Conserved tRNA-like Structural Motif in the Group I Intron Catalytic Core. *Cell* **1996**, *87* (6), 1135−1145.
(89) Buchmueller, K.L.; Weeks, K.M. Near Native Structure in an RNA Collapsed State. *Biochemistry* **2003**, *42* (47), 13869−13878.
(90) Sarrazin-Gendron, R.; Yao, H.-T.; Reinharz, V.; Oliver, C. G.; Ponty, Y.; Waldispühl, J.Stochastic Sampling of Structural Contexts Improves the Scalability and Accuracy of RNA 3D Module Identification*bioRxiv***2020**
(91) Loyer, G.; Reinharz, V. Concurrent prediction of RNA secondary structures with pseudoknots and local 3D motifs in an integer programming framework. *Bioinformatics* **2024**, *40* (2), btae022.
(92) Cruz, J. A.; Westhof, E. Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods* **2011**, *8* (6), 513−521.
(93) Antczak, M.; Zablocki, M.; Zok, T.; Rybarczyk, A.; Blazewicz, J.; Szachniuk, M. RNAvista: a webserver to assess RNA secondary structures with non-canonical base pairs. *Bioinformatics* **2019**, *35* (1), 152−155.
(94) Kwon, D. RNA function follows form − why is it so hard to predict? *Nature* **2025**, *639* (8056), 1106−1108.

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(95) Xu, B.; Zhu, Y.; Cao, C.; Chen, H.; Jin, Q.; Li, G.; Ma, J.; Yang, L. S.; Zhao, J.; Zhu, J.; et al. Recent advances in RNA structurome. *Sci. China Life Sci.* **2022**, *65* (7), 1285−1324.
(96) Leontis, B. N.; Lescoute, A.; Westhof, E. The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.* **2006**, *16* (3), 279−287.
(97) Rybarczyk, A.; Szostak, N.; Antczak, M.; Zok, T.; Popenda, M.; Adamiak, R.; Blazewicz, J.; Szachniuk, M. New in silico approach to assessing RNA secondary structures with non-canonical base pairs. *BMC Bioinf.* **2015**, *16* (1), 276.
(98) Watkins, M. A.; Rangan, R.; Das, R. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* **2020**, *28* (8), 963−976.
(99) Rivas, E.; Lang, R.; Eddy, R. S. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **2012**, *18* (2), 193−212.
(100) Sun, S.; Yang, J.; Zhang, Z. RNALigands: a database and web server for RNA-ligand interactions. *RNA* **2022**, *28* (2), 115−122.
(101) Su, H.; Peng, Z.; Yang, J. Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics* **2021**, *37* (1), 36−42.
(102) Panei, P. F.; Gkeka, P.; Bonomi, M. Identifying small-molecules binding sites in RNA conformational ensembles with SHAMAN. *Nat. Commun.* **2024**, *15* (1), 5725.
(103) Feng, Y.; Zhang, K.; Wu, Q.; Huang, S.-Y. NLDock: a Fast Nucleic Acid−Ligand Docking Algorithm for Modeling RNA/DNA−Ligand Complexes. *J. Chem. Inf. Model.* **2021**, *61* (9), 4771−4782.
(104) Sun, L.-Z.; Jiang, Y.; Zhou, Y.; Chen, S.-J. RLDOCK: A New Method for Predicting RNA−Ligand Interactions. *J. Chem. Theory Comput.* **2020**, *16* (11), 7173−7183.
(105) Mir, S.; Alhroub, Y.; Anyango, S.; Armstrong, D. R.; Berrisford, J. M.; Clark, A. R.; Conroy, M. J.; Dana, J. M.; Deshpande, M.; Gupta, D.; et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.* **2018**, *46* (D1), D486−D492.
(106) Sato, K.; Kato, Y. Prediction of RNA secondary structure including pseudoknots for long sequences. *Briefings Bioinf.* **2022**, *23* (1), 1−9.
(107) Seetin, G. M.; Mathews, H. D. RNA Structure Prediction: An Overview of Methods. *Methods Mol. Biol.* **2012**, *905*, 99−122.
(108) Kalvari, I.; Nawrocki, E. P.; Argasinska, J.; Quinones-Olvera, N.; Finn, R. D.; Bateman, A.; Petrov, A. I. Non-Coding RNA Analysis Using the Rfam Database. *Curr. Protoc. Bioinf.* **2018**, *62* (1), No. e51.
(109) Nawrocki, E. P.; Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29* (22), 2933−2935.
(110) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.
(111) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28* (23), 3150−3152.
(112) Rodrigues, J. P. G. L. M.; Teixeira, J. M. C.; Trellet, M.; Bonvin, A. M. J. J. pdb-tools: a swiss army knife for molecular structures. *F1000research* **2018**, *7*, 1961.
(113) Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.* **2007**, *56* (1−2), 215−252.
(114) Lu, X.-J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31* (17), 5108−5121.
(115) Lu, X.-J.; Bussemaker, H. J.; Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **2015**, *43*, No. e142.
(116) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422−1423.
(117) Johnson, P. Z.; Simon, A. E. RNAcanvas: interactive drawing and exploration of nucleic acid structures. *Nucleic Acids Res.* **2023**, *51* (W1), W501−W508.
(118) Mathews, D. H. How to benchmark RNA secondary structure prediction accuracy. *Methods* **2019**, *162*−*163*, 60−67.

https://doi.org/10.1021/acs.jcim.6c00108
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX