Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Inferring RNA structure from mobility-based deep mutational landscapes

**Yaoqi Zhou**

zhouyq@szbl.ac.cn

Shenzhen Bay Laboratory https://orcid.org/0000-0002-9958-5699

**Jinle Tang**

Institute of Physical and Systems Biology, Shenzhen Bay Laboratory

**Yazhou Shi**

Wuhan Textile University https://orcid.org/0000-0003-2579-3134

**Zhe Zhang**

Shenzhen Bay Laboratory

**Dailin Luo**

The Chinese University of Hong Kong

**Jian Zhan**

Shenzhen Bay Laboratory https://orcid.org/0000-0003-0856-2385

**Article**

**Keywords:**

**Posted Date:** March 6th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-8802544/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** There is **NO** Competing Interest.

# Inferring RNA structure from mobility-based deep mutational landscapes

Jinle Tang[1†], Yazhou Shi[1,2†], Zhe Zhang[1], Dailin Luo[1,3], Jian Zhan[1,4*], and Yaoqi Zhou[1*]

[1]Institute of Physical and Systems Biology, Shenzhen Bay Laboratory, Shenzhen, Guangdong Province, 518107, China

[2]Research Center of Nonlinear Science, School of Mathematics & Statistics, Wuhan Textile University, Wuhan, 430200, China

[3]School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, 518172, China

[4]Ribopeutic Inc., Guangzhou International Bio Island, Guangdong, 510320, China

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

* To whom correspondence should be addressed. Yaoqi Zhou. Tel: +86 755 26849275; Email: zhouyq@szbl.ac.cn; Jian Zhan. Tel: +86 755 26849280; Email: zhanjian@szbl.ac.cn.

## SUMMARY

Deciphering RNA three-dimensional structures remains a central challenge in molecular biology. While deep mutational profiling captures evolutionary constraints, its application has been limited to a few functions, such as ribozyme self-cleavage, that are compatible with high-throughput selection. Here we introduce RNA-MobiSeq, an integrative platform that leverages electrophoretic mobility as a proxy for native RNA structure. Native gel profiling of deep-mutational libraries enriches for structural homologs, and covariation analysis via an RNA-specific unsupervised method converts these data into effective folding restraints, enabling accurate prediction of base-pairing (*F1-score* ≥ 90%) and tertiary (3.6-11.2 Å RMSD) structures for nine diverse RNAs, including riboswitches, ribozymes, and long noncoding RNAs. Ranking as the best performer in CASP15-16 RNA targets predicted by the template-free techniques and validated through replicate and oligo-pool experiments, RNA-MobiSeq provides a general, probe-free framework for elucidating RNA structural landscapes, bridging sequence information and functional understanding.

## INTRODUCTION

Many RNAs, involved in fundamental biological processes of all living organisms, can self-assemble into specific base-pairing structures and rearrange their structures in response to other RNAs, proteins, and small molecules[1]. Despite decades of progress, determining high-resolution RNA structures remains a formidable challenge. Crystallization is often hindered by degenerate stable secondary structures and negatively charged phosphate backbones, whereas nuclear magnetic resonance (NMR) spectroscopy is limited by low number of intramolecular interactions and high conformational flexibility of RNA due to six torsion angles per residue[2,3]. Recent advances in cryo-electron microscopy (cryo-EM) have enabled structure determination of some RNAs and RNA–protein complexes within weeks, yet these successes are typically confined to large or highly ordered RNAs and demand sophisticated sample preparation and data analysis pipelines[4,5]. As a result, only about 3% of entries in the Protein Data Bank (https://www.rcsb.org/) contain RNA structures.

The limitation posed by high-resolution structure-determination techniques spurred the development of RNA-structure probing methods by employing enzymes, chemicals, hydroxyl radicals, and cross-linking techniques, more recently in combination with next-generation sequencing[6–8]. Multidimensional chemical mapping approaches, in particular, enable direct inference of base–base interactions[9]. One example is mutate-and-map, which infers RNA base pairs through combined sequence mutations with chemical mapping[10–12]. Another example is MOHCA, which employs highly reactive chemicals and the Fenton reaction to map the positions of nucleotides that are close together in three-dimensional structures by detecting damaged nucleotides near reacting Fe-modified nucleotide[9]. More recently, it has been shown that function-selected mutants from deep-mutational scanning alone can produce accurate base pairs through epistasis analysis[13,14] or analysis of covariation-induced deviation of activity (CODA)[15]. This technique of deep mutational scanning without chemical probing simplifies experimental requirements. Function-based selections, however, are limited to the RNAs with function amenable to high-throughput sequencing.

Parallel to experimental efforts, diverse computational approaches have been developed for RNA three-dimensional (3D) structure prediction. These methods can be categorized into homology modeling, fragment-based assembly, *de novo* folding (without using known structural fragments), and deep learning-based approaches[16-28]. Fragment assembly methods such as FARNA[16], MC-Sym[17], Vfold[18], 3dRNA[19], and RNAComposer[20], construct RNA tertiary models by combining structural fragments from template libraries followed by energy minimization. *De novo* folding methods attempt to fold RNAs into lowest-energy tertiary structures from unfolded states by using Monte Carlo (MC) or molecular dynamics simulations[21–23]. More recently, inspired by AlphaFold[24,25], some deep learning-based methods have emerged to extract the geometric structural information[25–28], which serves as restraints for constructing RNA tertiary structures using traditional energy-based methods or to make an end-to-end prediction such as AlphaFold 3. However, unlike the success of AlphaFold in high-precision protein-structure prediction[24], the overall performance for RNA structure-prediction techniques remains low[29–32].

The above persistent challenges highlight the need for experimental–computational strategies that are both broadly applicable and experimentally accessible. Prior studies have shown that single-nucleotide mutations can globally reshape RNA conformational ensembles[33-35], and destabilizing variants often exhibit altered electrophoretic mobility in native gels gels[35,36]. For instance, while wild-type n-Tr20 (tRNA) adopts a single conformation, its $C_{50}U$ mutant populates two distinct states[35]; similarly, the $C_{10320}G$ mutation in *DVxrRNA1* dramatically reduces gel mobility[36]. These observations suggest that the fractional retention of mutants in the native PAGE band could serve as a quantitative proxy for their structural "fitness" relative to the wild type.

Here, we integrate this concept with high-throughput sequencing and computational modeling to develop RNA-MobiSeq, an unsupervised framework for RNA structure inference. By deriving *native-mobility-fitness* scores from sequencing mutational libraries before and after mobility selection and applying RNA-specific CODA analysis (CODA2), we generate covariation restraints

that, combined with the BRiQ statistical energy function[23], produce high-accuracy secondary base-pairing structures (*F1-score* ≥ 0.9) and tertiary models (median LDDT ≥ 0.58, RMSD of 3.6–11.2 Å) across nine RNAs spanning ribozymes, riboswitches, and lncRNAs (55–205 nt). Validated through CASP benchmarks and replicate experiments, RNA-MobiSeq provides a robust, probe- and template-free route for elucidating RNA tertiary structural landscapes.

## RESULTS

### Overview of the RNA-MobiSeq experimental-computational platform

The RNA-MobiSeq platform integrates native gel-based selection with sequencing and computational modeling to infer both base-pairing and tertiary structures directly from RNA sequence (Fig. 1). The framework begins by generating artificial homologous sequences and quantifying their native-mobility fitness (a proxy for structural integrity) via high-throughput sequencing of RNA mutants selected by gel mobility. These quantitative mobility-derived fitness scores are then employed to identify covarying base pairs through an RNA-specific, unsupervised analysis, which subsequently guides energy-based 3D modeling. Briefly, error-prone PCR (EP-PCR) or oligonucleotide pool synthesis was utilized to construct mutational DNA libraries, which were transcribed *in vitro* into RNA. The resulting mutant pools were separated by native PAGE according to electrophoretic mobility, and the band corresponding to the wild-type mobility was excised for reverse transcription. Both pre- and post-selection samples were deep-sequenced to obtain mutant frequencies. The relative fraction of each variant in the excised band was employed to estimate its native-mobility fitness score, approximating its structural similarity to the wild type. These fitness scores were then analyzed using the RNA-specific CODA2 (COvariation-induced Deviation of Activity) algorithm and Monte Carlo simulated annealing to detect covariation patterns indicative of base-pairing interactions. The resulting secondary structure models were incorporated as restraints into 3D modeling using the BRiQ statistical potential and a folding-tree-based energy minimization algorithm[23]. Detailed protocols are provided in the Methods section.

To test this experimental-computational platform, we first chose four RNAs in three different types[37-41]: two riboswitches, an exonuclease-resistant RNA (xrRNA), and a self-cleaving ribozyme. They are SAM-VI riboswitch from *Bifidobacterium angulatum*[37] (PDB ID: 6LAS, sequences 1-55, 55 *nt*), adenine riboswitch from *Vibrio vulnificus*[38] (PDB ID: 5E54, sequences 13-83, 71 *nt*, exonuclease resistant RNA from *Zika* virus[39] (PDB ID: 5TPY, sequences 1-71, 71 *nt*), and inactive ribozyme mutant CPEB3-$C_{57}U$[40] (81 *nt*) from human (PDB ID: 7QR4[41]). Their base-pair structures and 3D structures are shown in Figs. 2A, B, C, and D, respectively. These RNAs were chosen for their relatively small sizes and known stable three-dimensional structures. Two riboswitches were chosen to compare the effect of base pairs with or without tertiary-interaction-stabilized pseudoknots for adenine riboswitch and SAM-VI riboswitch, respectively, on mobility selection. CPEB3 ribozyme was chosen so that one can compare mobility-based selection to function-based selection conducted previously[15]. In addition, we employed four RNA targets from CASP 15 and CASP 16[29,30], each exceeding 120 *nt* in length. Two CASP 15 targets classified as "difficult", which had no published homologous structures before the challenge[29], were included: R1116, a 157 *nt* Cloverleaf RNA (PDB ID: 8S95), and R1149, a 124 *nt* SARS-CoV-2 SL5 (PDB ID: 8UYS), and their structures were solved by X-ray crystallography and cryo-EM, respectively[42,43]. From CASP 16, we selected R1203, a 134 *nt* HIV-1 Rev Response Element stem-loop II (RRE SLII, PDB ID: 8UO6 [44], classified as "medium" difficulty, solved by X-ray crystallography), and R1242, a 205 *nt* raiA ncRNA motif (raiA, PDB ID: 9ELY[5,45], classified as "difficult", solved by cryo-EM)[30].Their distinct base-pairing structures and 3D structures are shown in Figs. 2E-2H, respectively. All eight structures are chosen very different from each other (structural similarities between them are < 0.11 according to TM-Score, Supplementary Table S1). To further assess its performance in blind prediction, we also tested RNA-MobiSeq on three recent RNA-Puzzles targets (https://www.rnapuzzles.org/open-puzzle/): PZ50 (57 nt), PZ51 (114 nt), and PZ62 (91 nt). While the reference structure of PZ50 has been released (PDB ID: 8JY0)[46], those of PZ51 and PZ62 remain

undisclosed at the time of analysis.

**Mobility-based Selection**

We first tested if mutations could lead to a visible change in mobility on a native PAGE for six RNAs studied including two CASP 15 targets as described above. For each RNA, we randomly introduced two mutations that will disrupt two base pairs in two different stems. The chosen double mutations were shown in Figs. 2A-2F for the six RNAs, respectively (all RNA sequences are listed in Supplementary Table S2). Figs. 2I-2N compare the mobilities of wild type (WT), a disruptive double mutant (2M), and two RNA mutant-library replicates generated from EP-PCR. In five of the six RNAs, the disruptive mutations led to a clear change in mobility relative to the WT, indicating structural perturbation. The remaining RNA, SARS-CoV-2 SL5, exhibited only a subtle shift (Fig. 2N), but the mutant band displayed weaker intensity, indicating potential alterations in the native structure. In all cases, mutant libraries displayed broader bands than the corresponding WT. Thus, the native PAGE result confirmed that mobility is sensitive enough to separate at least some structure-disrupting mutations from structure-unaltered mutations. Three destabilized RNA mutants (2M: adenine riboswitch, xrRNA, and CPEB3 ribozyme) led to slower mobility compared to the wild type (upper in the figure), which is expected for loosened structural forms with a larger hydrodynamic radius. However, the opposite is true for SAM-VI riboswitch and Cloverleaf RNA. This phenomenon has also been reported in other studies. For example, mutants of n-Tr20$C_{50}$U, n-Tr20$G_{64}$A, and n-Tr20$C_{50}$U/$G_{64}$A migrate faster than their wild type[35]. The reason for this peculiar behavior is not entirely clear but may be related to higher charge densities caused by structural changes.

For mobility-based selection, the gel band corresponding to the wild-type mobility was excised, under the assumption that RNAs with similar mobilities adopt similar structural ensembles. Similarly, mutation libraries and corresponding bands were also prepared for two CASP 16 targets (Figs. 2G,

2H, 2O and 2P) and three RNA puzzle targets (Figs. S1A-C), without testing individual double mutation variants.

## Quality of the Mutation Libraries

For covariation analysis, the most informative variants are single and double mutants. We therefore optimized mutation rates to enrich RNA libraries for these categories. High-throughput sequencing of pre-selection libraries confirmed 100% coverage of all single mutants and >82% coverage of double mutants across eleven RNAs (95.2%, 88.4%, 94.4%, 91.7%, 82.3%, 92.0%, 87.8%, 81%, 97.8%, 91.2%, and 91.2% for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, raiA, PZ50, PZ51, and PZ60, respectively; Supplementary Table S3). This high double-mutation coverage ensures sufficient statistical power for subsequent covariation analysis. EP-PCR procedures can be simplified by using oligo-pools as demonstrated for two CASP 15 targets and discussed in the discussion section. For each mutant, we computed its native-mobility-fitness score as the proportion of reads recovered in the gel band corresponding to the wild-type mobility, normalized by the wild-type pre-/post-selection ratio (See Methods).

## RNA-specific Co-variation Analysis & Signal Amplification (2D folding)

To significantly improve the sensitivity of detecting co-variation, we developed an RNA-specific technique named CODA2. Previously, CODA provided unsupervised detection of base pairs by searching for covariant outliers (the actual fitness score of a double mutant is significantly more than the predicted fitness score of the double mutant from the fitness values of the two corresponding single mutant variants through Support Vector Regression, SVR)[15]. Unlike CODA, which utilized fixed SVR hyperparameters for all RNAs, CODA2 improves sensitivity by minimizing base-pairing energy through a grid search of the SVR parameter space for each RNA (i.e., RNA-specific C and $\gamma$ in the SVR model, as well as the variance truncation coefficient $n$) (Fig. S2, see Methods for details).

Moreover, CODA2 incorporates mutation positions and base types before and after the mutations as the input features, in addition to two single-mutation fitness values in CODA. These refinements markedly improved the detection of base-pairing signals across eight RNAs (Fig. 2Q). The *F1*-scores (harmonic mean of precision and sensitivity with 1, or 100%, as the highest value) achieved by CODA2 were 0.34, 0.54, 0.64, 0.39, 0.42, 0.26, 0.25, and 0.56 for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA, which improved over CODA by 89%, 69%, 31%, 70%, 121%, 100%, 56%, and 60%, respectively. Consistent with the *F1-score* trends, Matthews Correlation Coefficients (*MCC*s) also confirmed the enhanced predictive accuracy of CODA2 (Supplementary Table S4).

Fig. 2Q also presents a comparison of CODA2's performance in inferring true base pairs from deep mutation data of the eight RNAs with other state-of-the-art covariation-analysis methods including R-scape[47], mean-field direct coupling analysis (mfDCA-RNA)[48], pseudo-likelihood maximization coupling analysis (plmDCA-RNA)[49], and epistasis analysis from Schmiedel & Lehner (epistasis-SL)[13] as well as from Rollins *et al* (epistasis-Rollins)[14]. Here, the thresholds for defining base pairs were all chosen to maximize the *F1-score* to facilitate comparison. The four methods (i.e., mfDCA-RNA, plmDCA-RNA, epistasis-SL, and epistasis-Rollins) have similar performance in terms of *F1-score* for SAM-VI riboswitch (*F1-score* $\leq 0.14$), adenine riboswitch (*F1-score* $\leq 0.21$), xrRNA (*F1-score* $\leq 0.19$), CPEB3 ribozyme (*F1-score* $\leq 0.13$), Cloverleaf RNA (*F1-score* $\leq 0.11$), SARS-CoV-2 SL5 (*F1-score* $\leq 0.10$), RRE SLII (*F1-score* $\leq 0.12$), and raiA (*F1-score* $\leq 0.22$). The best *F1-scores* of base-pairs inferred from R-scape (0.10, 0.31, 0.40, 0.21, 0.15, 0.12, 0.16, and 0.28 for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA, respectively) are mostly better than traditional covariation and epistasis analysis (except in the case of SAM-VI riboswitch). Its performance, however, is behind the second-best CODA (Fig. 2Q), whereas CODA2 has an average improvement of 74.5% over CODA and 92.7% over R-scape for 8 RNAs studied here.

Covariation signals captured by CODA2 can be amplified by MC simulated annealing for 2D folding (CODA2 + MC) with commonly employed experimental energetic parameters, and subsequent refinement through folding by BRiQ with the restraints from CODA2 + MC analysis (CODA2 + MC + BRiQ). Across 1000 independent MC simulated annealing runs and 5000 BRiQ folding simulations, the top 100 structures were utilized to produce a distribution of pairing results. Fig. 2Q shows that CODA2 + MC (best in the top 5) yields an increase in *F1-score* over CODA2 by 150%, 65%, 42%, 121%, 117%, 269%, 252%, and 61% for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA respectively. The *F1-scores* for all RNAs ≥ 0.85 after MC simulated annealing and are then further improved to ≥ 0.90 after BRiQ refinement, indicating a high-accuracy base-pair prediction for all eight RNAs (Fig. 2Q; Fig. S3).

**Capturing Tertiary Base Pairs**

To evaluate whether mobility-derived covariation signals extend beyond local secondary-structure and canonical base pairing, we analyzed the recovery of tertiary interactions across eight RNAs (Fig. 3; Fig. S3). Tertiary base pairs were defined as pseudoknots, lone base pairs, noncanonical base pairs, and base triplets. SARS-CoV-2 SL5 is the only one that does not have any tertiary base pairs. SAM-VI riboswitch has two noncanonical base pairs. Adenine riboswitch has five pseudoknotted base pairs including two noncanonical base pairs. xrRNA has six pseudoknotted base pairs including two lone base pairs and two noncanonical base pairs. CPEB3 ribozyme has seven pseudoknotted base pairs including one noncanonical base pair and two additional noncanonical base pairs. Cloverleaf RNA has one lone pseudoknotted base pair, three noncanonical base pairs, and two triplets. RRE SLII contains four lone pseudoknotted base pairs and two noncanonical base pairs, and raiA features two pseudoknots, comprising six and two consecutive base pairs, respectively.

As shown in Fig. 3 and Fig. S3, CODA2 alone effectively captured pseudoknot signals in the

adenine riboswitch, xrRNA, CPEB3 ribozyme, RRE SLII, and raiA. Although a tertiary signal for the lone pseudoknot in Cloverleaf RNA was present (Fig. 3M), it was not strong enough to persist among the top five predictions after MC and BRiQ refinement. Signals for noncanonical base pairs and triplets as well as distance-based tertiary interactions were also difficult to capture. Nevertheless, folding by the BRiQ statistical energy function can improve base pairs from CODA2 + MC by removing unphysical false positives and including some distance-based contacts in SAM-VI riboswitch ($U_{35}$•$G_{50}$), adenine riboswitch ($A_{12}$•$C_{41}$), xrRNA ($U_4$•$C_{23}$), CPEB3 ($A_5$•$G_{63}$/$C_{34}$•$U_{66}$), Cloverleaf RNA ($C_{41}$•$C_{76}$), SARS-CoV-2 SL5 ($C_{25}$•$A_{95}$), RRE SLII ($G_{19}$•$G_{111}$), and raiA ($G_{39}$•$A_{132}$/$C_{89}$•$G_{154}$) (Figs. 3C, 3F, 3I, 3L 3O, 3R, 3U, and 3X), as well as some noncanonical base pairs located within stems or hairpin loops (e.g., $U_{86}$-$U_{105}$, $C_{87}$-$U_{104}$, and $C_{88}$-$C_{103}$ in Cloverleaf RNA; Fig. S3E). Notably, CODA2 also revealed extensive long-range base pairing in the 205-nt raiA lncRNA motif[5,45], capturing contacts bridging its 5' and 3' ends (e.g., $G_7$-$C_{202}$, $U_9$-$A_{200}$, $A_{10}$-$U_{199}$, and $G_{11}$-$C_{198}$), alongside two pseudoknotted elements (e.g., $G_{39}$-$C_{89}$, $C_{40}$-$G_{88}$, and $A_{41}$-$U_{87}$ in one pseudoknot, and $U_{84}$-$A_{184}$ in the second) (Fig. S3). Although the signal for the second pseudoknot was weaker due to its short two-pair stem, subsequent Monte Carlo refinement and BRiQ optimization retained these interactions, yielding an overall *F1-score* of 0.91 for the best in top 5.

In fact, across all eight RNAs with sequence lengths ranging from 55 to 205 nt, we achieved *F1-scores* $\geq$ 0.9 for base-pair inference (Fig. S3), despite the diverse shapes and functions of these RNAs (riboswitches, ribozymes, viral RNA elements, and long non-coding RNAs). These results demonstrate that mobility-informed covariation analysis not only captures canonical base pairs with high fidelity but also retains the capacity to recover complex tertiary architectures underlying RNA folding and function.

**Comparison with Other Experimental Base-Pair Mapping Methods**

Other experimental techniques for mapping base pairs were applied to Adenine riboswitch

[Mutation-and-map by SHAPE (selective 2'-hydroxyl acylation and primer extension)[10]], xrRNA [M2-Seq by DMS (dimethyl sulfate) probe[12]], and SARS-CoV-2 [M2-Seq by DMS probe[43]] (Fig. S4). As summarized in Table 1, CODA2 inference from RNA-MobiSeq outperformed the Mutate-and-map (SHAPE) on adenine riboswitch, M2-Seq (DMS probe) on xrRNA, and M2-Seq (DMS probe + Large library) on SARS-CoV-2 SL5, improving *F1-scores* by 8%, 220%, and 86%, respectively. For SARS-Cov-2 SL5, the optimized M2-Seq protocol (DMS probe + Scarless) captures more structural information (*F1-score* = 0.62) than CODA2 (*F1 score* = 0.26) but CODA 2 + MC recovered the weak signals to yield an *F1-score* of 0.96 (Table 1). The enhanced performance of the "Scarless" protocol was attributed to its exclusion of potential secondary-structure interference caused by the primer binding sequence during DMS probing[43]. The same technique, despite added experimental complexity and manual labor, may well be employed for improving RNA-MobiSeq. It should be noted that compared to M2-Seq, the experimental setup of mobility-based selection is simpler as it does not utilize chemical probes.

**3D Structure Prediction**

To assess how mobility-informed covariation contributes to tertiary structure formation, we evaluated the accuracy of 3D models generated by RNA-MobiSeq using the BRiQ statistical potential. Structural similarity was quantified by the root-mean-square deviation (RMSD) and local distance difference test (LDDT) between the predicted 3D structures and the corresponding native structure for each RNA. The number of all types of steric clashes per thousand residues (clash score) was also calculated.

As shown in Figs. 4A-H, RNA-MobiSeq accurately recapitulated the global architectures of eight RNAs, with top 1 predicted structures (according to the energy function) showing RMSDs of 7.3 Å, 7.1 Å, 4.8 Å, 6.0 Å, 8.7 Å, 8.8 Å, 8.8 Å, and 13.8 Å and LDDTs of 0.70, 0.64, 0.71, 0.61, 0.59, 0.66, 0.70, and 0.52 for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf

RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA, respectively. The best of the top 5 predictions further improved to RMSDs of 6.1 Å, 5.9 Å, 3.6 Å, 5.2 Å, 7.7 Å, 5.1 Å, 7.5 Å, and 11.2 Å and LDDTs of 0.7, 0.66, 0.74, 0.65, 0.61, 0.73, 0.72, and 0.58. Overall structural topology is well captured for each of the eight RNAs (best in Top 5 all with ≥ 0.58 LDDT, RMSD ≤ 11.2Å). Consistently, Fig. S5 shows that lower BRiQ energy scores correlate with lower RMSDs across all eight RNAs.

**3D Structure Prediction Compared to BRiQ Folding with RNAfold-Derived Secondary Structures as Restraints.**

To obtain a sense of importance in tertiary interactions obtained by RNA-MobiSeq, we compared secondary (in *F1-score*s or distribution of *F1-score*s) and tertiary structures (in RMSD, LDDT, and clash scores) for the eight RNAs by using the commonly-used secondary-structure-predictor RNAfold[50], instead of CODA2 + MC as restraints for BRiQ folding.

For RNAs containing pseudoknots, including the adenine riboswitch, xrRNA, CPEB3 ribozyme, and raiA, RNA-MobiSeq substantially outperformed RNAfold-based folding. The best in top 5 models achieved RMSD improvements of 178%, 252%, 292%, and 150%, respectively (Figs. 5B-D and 5H; Fig. S6). This large improvement is somewhat expected because RNAfold cannot predict pseudoknots. Indeed, the *F1-score*s for base pairs of the best of the top 5 predictions obtained from CODA2 + MC are 0.89, 0.91, 0.86, and 0.91 for adenine riboswitch, xrRNA, CPEB3 ribozyme, and raiA, respectively, which are 11%, 68%, 85%, and 184% higher than those from RNAfold. Notably, the performance of CODA2 itself may not be better than RNAfold (better for two RNAs but worse for the other six RNAs), as it identifies only strongly covarying base pairs; however, enhancing the signals by MC simulated annealing (2D folding) improves over RNAfold for those RNAs with pseudoknot stems.

Interestingly, the improvement is even observed for SAM-VI riboswitch, Cloverleaf RNA,

SARS-CoV-2 SL5, and RRE SLII in 3D structure prediction, despite these RNAs do not have or have few lone tertiary base pairs, and CODA2 + MC performs comparable to RNAfold (Figs. 5A, 5E, 5F, and 5G) in predicted base pairs. The top 1 predicted structure for SAM-VI riboswitch is 7.3 Å RMSD and 0.70 LDDT by RNA-MobiSeq, compared to 7.5 Å RMSD and 0.62 LDDT by RNAfold + BRiQ. The difference is even larger for Cloverleaf RNA (8.7 Å RMSD by RNA-MobiSeq versus 12 Å RMSD by RNAfold + BRiQ), SARS-CoV-2 SL5 (8.8 Å versus 10.5 Å), and RRE SLII (8.8 Å versus 15.0 Å). These improvements occurred despite similar or worse performance in base pair prediction by CODA2 + MC for SAM-VI riboswitch (*F1 score* of 0.89 by RNAfold versus 0.77 by CODA2 + MC (top 1)), Cloverleaf RNA (0.96 versus 0.94), SARS-CoV-2 SL5 (0.95 versus 0.96), and RRE SLII (0.92 versus 0.78), suggesting that the pairing probabilities from CODA2 may have contained tertiary distance information. In the case of Cloverleaf RNA, while CODA2 + MC failed to precisely predict the lone pseudoknotted base pair ($G_{20}$-$C_{137}$) and base triplets ($A_{70}$•$C_{87}$-$U_{104}$, and $G_{77}$•$G_{82}$-$C_{109}$) essential for the stabilization of its 3D structure, the predicted nearby false-positive base pairs (e.g., $G_{19}$-$C_{129}$, $G_{20}$-$C_{128}$, $C_{69}$-$G_{95}$, and $A_{70}$-$U_{94}$; see Fig. S3E) may have nonetheless promoted the formation of native-like topologies with significantly lower RMSD values. A similar phenomenon was observed for RRE SLII, where predicted base pairs $G_{18}$–$C_{114}$ and $G_{19}$–$U_{113}$, though incorrect, closely approximate the native pseudoknotted interactions $G_{18}$–$C_{113}$ and $G_{19}$–$U_{114}$.

## 3D Structure Prediction in Comparison to Other Structure Prediction Methods

For four small RNAs (SAM-VI riboswitch, adenine riboswitch, xrRNA, and CPEB3 ribozyme), we compared RNA-MobiSeq with the current state-of-the-art RNA 3D structure prediction methods, including four deep-learning-based (AlphaFold 3[25], trRosettaRNA[26], DRfold[27], and Rhofold[28]) and three energy-function-based (MC-Fold/MC-Sym[17], Vfold[18], and SimRNA[22]) approaches; see in Methods. As shown in Fig. 4I, RNA-MobiSeq outperformed all the energy function-based methods in RMSD values (the only method with RMSD ≤ 6.1 Å for all 4 RNAs). For deep-learning-based methods, comparisons were only made for CPEB3 ribozyme because the other three structures are in

the training data for these methods[25–28]. For CPEB3, the best model among the top 5 RNA-MobiSeq predictions achieved an RMSD of 5.2 Å, representing a 29% improvement over the second-best result (7.3 Å from trRosettaRNA).

For four CASP RNAs (Cloverleaf RNA and SARS-CoV-2 SL5 from CASP 15, and RRE SLII and raiA from CASP 16), we compared RNA-MobiSeq with predictions reported by CASP methods (https://predictioncenter.org/index.cgi)[29,30] as well as AlphaFold 3. Because RNA-MobiSeq does not rely on known homologous sequences or 3D templates, CASP submissions that explicitly utilized structural templates or integrated multiple 3D structure prediction pipelines were excluded to ensure a methodologically consistent comparison; see in Methods. Figure 4J shows the subset of predictions with RMSD < 20 Å used in this analysis, in which RNA-MobiSeq achieved the lowest RMSD (best among the top 5 predictions) for all four CASP targets. For target R1116 (Cloverleaf RNA), the top 3 models among the 5 RNA-MobiSeq predictions all outperformed the top-ranked prediction from the second-best method (Yang-Server) with RMSD of 10.9 Å, which is 42% larger than RNA-MobiSeq top-1 model (7.7 Å). For target R1149 (SARS-CoV-2 SL5), RNA-MobiSeq obtained the best result (5.1 Å RMSD), surpassing the leading performance in CASP 15 by approximately 26% (6.9 Å from DF_RNA). For the CASP16 targets R1203 (RRE SLII) and R1242 (raiA), the best models among the top 5 RNA-MobiSeq predictions achieved RMSDs of 7.5 Å and 11.2 Å, corresponding to improvements of 11% and 2%, respectively, over the second-best AlphaFold3-server predictions.

Collectively, these comparisons highlight RNA-MobiSeq as a unsupervised template/fragment-free framework that rivals or exceeds leading AI- and energy-based methods, underscoring the power of integrating deep mutational scanning, mobility selection, CODA2, MC simulations, and BRiQ modeling for accurate RNA base pairing and 3D structure inference from sequences.

**Evaluation of RNA-MobiSeq on RNA-Puzzles Targets**

To further assess the generalizability and blind-test performance of RNA-MobiSeq, we applied the

RNA-MobiSeq platform to three recent RNA-Puzzles targets: PZ50, PZ51, and PZ62. Among them, only PZ50 has a publicly available native structure (RhoBAST, PDB: 8JY0)[46] after our experiments and structure prediction completed, whereas PZ51 and PZ62 remain blind at the time of submitting this paper. For PZ50, the CODA2-inferred contacts achieved an *F1-score* of 0.60, successfully capturing signals for all three major stems, as well as two notable lone tertiary base pairs ($G_{23}$–$C_{44}$ and $G_{27}$–$C_{35}$); see Fig. S1A. Although the inclusion of MC slightly reduced the prominence of these tertiary interactions due to the penalization of lone base pairs, the resulting base-pairings (e.g., the best in top 5 with *F1-score* = 0.89) were sufficiently accurate to serve as input constraints for 3D structure modeling using BRiQ, yielding an *F1-score* of 0.91 for the best 3D model in top 5. The final predicted tertiary structures reproduced the overall topology of the RNA, with the top 1 predicted structure reaching RMSD = 6.6 Å, LDDT = 0.55, and the best in top 5 achieving RMSD = 5.3 Å, LDDT = 0.58. This accurate prediction was achieved despite of likely structural differences between the target (pure RNA) and the experimental structure (protein-RNA-ligand complex). In the solved RhoBAST structure, the 5'-hairpin loop is longer than in the PZ50 target and interacts with a protein, bringing it closer to the 5'-stem. Additionally, the 3'-hairpin loop binds a ligand (TMR-DN) and interacts more extensively with the junction loop, resulting in a more compact tertiary structure[46].

For PZ51 and PZ62 (Figs. S1B and C), CODA2 revealed distinct covariation patterns consistent with unique structural organizations. PZ51 appears to form a three-way junction without evident tertiary contacts, whereas PZ62 exhibits a more intricate topology comprising five stems and a pronounced pseudoknot. The top 5 predicted 3D structures for both targets have been submitted to a public repository (see Data and Code Availability) and will be evaluated upon release of the native structures as part of the ongoing blind assessment.

# DISCUSSION

RNA structures are challenging to determine experimentally and predict computationally. Here, we propose a simple and yet effective method called RNA-MobiSeq that couples an age-old molecular biology technique (mobility shifts in native PAGE) with modern high-throughput sequencing to produce artificial homologous sequences with fitness scores, which permit extraction of covariant base pairs by a sensitive RNA-specific CODA algorithm, followed by MC simulated annealing and base-pair-restrained structure prediction guided by the BRiQ energy function.

It is established that mutation-induced structural changes can change RNA mobilities in electrophoresis[35,36]. What is not known is whether mobility can serve as a tool to separate structural homologs from structurally destabilized mutants in an RNA mutant library for effective covariation analysis. Cutting the band with the same mobility as the wild type provides the fraction of mutants that maintain a tertiary structure similar to the wild type in their conformational ensemble. Although the gel experiment may not be sensitive enough to distinguish all mutation-induced structural changes, e.g. structural rearrangement, and could introduce certain systematic noises or random errors during manual cutting, the co-variation signals remain strong enough to allow the extraction of secondary, all pseudoknot stems, and some tertiary contacts by CODA2, MC simulated annealing, and BRiQ-based optimization for nine RNAs with known structures (Figs. 3A, 3D, 3G, 3J, 3M, 3P, 3S, 3V; Fig. S1A), achieving high-accuracy secondary structure predictions at the CODA2 + MC level and beyond (Fig. 2Q; Fig. S1).

In native PAGE, PZ62 showed two closely spaced bands, while the other ten wild-type RNAs each showed a single major band (Fig. 2I-P; Figs. S1 and S7). Both bands in PZ62 were excised separately and used as references for library screening. CODA2 analysis revealed that one of the two bands corresponded to a well-defined secondary structure (Fig. S1C), whereas the other exhibited disordered signals with no apparent stem formation (Fig. S7), indicating that subtle mobility

differences can reflect distinct structural states. Overall, the mutant libraries at each wild type position generally displayed lighter and more diffuse bands, along with the appearance of some mutants at other mobilities (Figs. 2I-P, and Fig. S1), confirming that electrophoretic mobility is a good indicator of structural integrity.

To assess the reproducibility of RNA-MobiSeq, we performed replicate experiments for two representative RNAs (i.e., xrRNA, and CPEB3 ribozyme) under the same experimental conditions as the original runs (Fig. S8). Replicates exhibited highly consistent mobility shift patterns, with mutant libraries showing broader and weaker bands than their wild-type counterparts. CODA2-based base-pairing inference yielded *F1-scores* of 0.58 and 0.37 closely matching the original values of 0.64 and 0.39 for xrRNA and CPEB3 ribozyme, respectively, reflecting relative deviations below 10%. Further validation was performed by conducting MC folding simulations using the replicate CODA2 scores for xrRNA and CPEB3 ribozyme. The *F1-scores* (best in top 5) were 0.89 and 0.87, nearly identical to the original values of 0.91 and 0.86, respectively. These similar base-pairing patterns also translated into consistent 3D structure predictions using BRiQ: the best RMSDs (in top 5) were 3.9 Å for xrRNA and 5.2 Å for CPEB3 ribozyme in the replicates, compared to 3.6 Å and 5.2 Å in the original predictions. These findings demonstrate the high reproducibility of RNA-MobiSeq and reinforce its utility as a reliable approach for RNA base-pairing and 3D structure inference.

It is of interest to compare mobility-based to function-based selection done previously for CPEB3 ribozyme[15]. The double mutation coverage was only 41% by function selection, compared to 92% in this work. Despite the poorer coverage, function-based selection leads to a much higher performance of base pair inference (0.55 versus 0.24 for *F1-score* by CODA). The *F1-score* of CPEB3 ribozyme for mobility-based screening reaches 0.39 only after using a more sensitive CODA2, which is still far below 0.55. The band for the CPEB3 ribozyme mutant library is quite faint and of lower quality compared to the function-selected library[15] (Fig. 2J), indicating that most mutants negatively

impacted structural stability. Nevertheless, enough RNA mutants (single and double mutants) were extracted (see Supplementary Table S3), so that the key pseudoknot in CPEB3 is captured by mobility selection and high-accuracy prediction of base pairs was obtained after MC simulated annealing (*F1-score* = 0.86), as shown in Fig. S3D. This ultimately leads to the predicted CPEB3 ribozyme structure by RNA-MobiSeq that is substantially more accurate than those produced by AlphaFold 3, trRosetta, DRfold, RhoFold, Vfold, SimRNA, and MC-Sym. It is the only one with an RMSD ≤ 6 Å (Fig. 4I). Thus, although mobility selection is not as clean as selection by self-cleavage, CODA2 + MC can rescue. More importantly, unlike function selection, mobility-based selection can apply to all structured RNAs, independent of function. Consistent high performance across nine RNAs with different functions (*F1-score* for base pair prediction: ≥ 0.85 by CODA2 + MC and ≥ 0.9 after BRiQ folding; RMSD for 3D structure prediction: 3.6-11.2 Å for the best in the top 5) underscores RNA-MobiSeq as a generalizable and reliable framework for structural inference.

One potential improvement is to utilize oligo pools rather than EP-PCR in RNA-MobiSeq. This would significantly shorten experimental times as controlling the mutation rate by EP-PCR requires trial and error. We performed RNA-MobiSeq for CASP 15 RNA targets R1116 (Cloverleaf RNA) and R1149 (SARS-CoV-2 SL5) by synthesizing all single and double mutants as oligo pools (Fig. S9A). Although the coverage for double mutations from oligo pools is higher than that from EP-PCR (i.e., 93.3% versus 82.3% for R1116, and 99.5% versus 92.0% for R1149; see Supplementary Table S3), the *F1-scores* by CODA2 from the sequences generated with RNA-MobiSeq-OP are not always superior to those from EP-PCR (0.28 versus 0.42 for R1116, and 0.32 versus 0.26 for R1149, respectively; see Figs. 3M & 3P; Fig. S9). For R1149, while there was an improvement in *F1-score* given by CODA2, the improvement did not lead to a prediction of more accurate structures because of similar base-pair accuracy after MC simulated annealing (0.96 versus 0.96 for the *F1-score* given by CODA2 + MC for oligo-pool-based and EP-PCR-based RNA-MobiSeq, respectively); see Fig. S9D. Thus, only the EP-PCR result was reported in the result section.

The usefulness of mobility-selected structural homologs was further demonstrated by comparison with RNAfold + BRiQ and AIChemy_RNA2. Although RNAfold + BRiQ employed the same BRiQ energy function as RNA-MobiSeq, the restraints yielded from mobility-selected structural homologs are consistently better than RNAfold, not only for pseudoknot-containing RNAs (Figs. 5B-5D & 5H), but also for these RNAs with few tertiary base pairs (Figs. 5A & 5E-5G), suggesting that mobility-selected structural homologs contain more useful tertiary structural information. Furthermore, large improvement of RNA-MobiSeq over AIChemy_RNA2 in two CASP 15 targets (RMSD reduced by > 51%), despite the latter utilizing chemical probing data or homology templates as folding restraints and the same BRiQ potential[29], indicates the consistent quality data from mobility-selected structural homologs.

We also examined whether base-pair restraints inferred from CODA2 + MC could enhance other folding techniques. We tested it on MC-Sym, Vfold, and SimRNA on SAM-VI riboswitch, adenine riboswitch, xrRNA, and CPEB3 ribozyme. MC-Sym did not yield results. Vfold yielded the result for adenine riboswitch with RMSD changed from 6.9 Å to 6.4 Å. Although the Vfold pipeline did not predict pseudoknotted base pairs in adenine riboswitch, the accurately predicted secondary structure (*F1-score* = 0.83), combined with the existing homology riboswitch junction loop templates, allowed it to assemble a native-like structure[18]. SimRNA yielded results for the four RNAs with significant improvement for RMSD changed from 23.4 Å, 23.3 Å, 16.5 Å, and 23.3 Å to 9.2 Å, 11.6 Å, 12.6 Å, and 10.7 Å for SAM-VI riboswitch, adenine riboswitch, xrRNA, and CPEB3 ribozyme, respectively (Fig. S10). Thus, base pairs inferred from mobility-selected structural homologs are useful for improving other folding-based methods.

The BRiQ potential itself, while one of the most effective statistical energy functions for RNA (enabling AIChem_RNA2 to achieve the #1 in CASP 15[23,29]), still exhibits intrinsic limitations. When applied to refine native RNA structures, BRiQ failed to produce consistently lower energies than those of CODA2 + MC + BRiQ-predicted models (Fig. S5), suggesting that native

conformations are not always located at the global minima of the BRiQ landscape. BRiQ's Monte Carlo simulated annealing algorithm may also suffer from insufficient sampling when folding larger RNAs. Together, these results suggest that RNA-MobiSeq's inability to recover high-accuracy, near-native structures (< 2 Å RMSD) stems from both limitations in the BRiQ energy function and incomplete conformational sampling rather than deficiencies in the experimental covariation signals. For proteins, AlphaFold2 successfully overcomes this energy-function problem by performing end-to-end training[24,25]. Such AI-based advances are yet to achieve for RNA structure prediction due to the lack of a sufficient number of RNA structures[32].

Recent results from CASP 16 highlight that RNA 3D structure prediction remains largely in the "template-based" phase, with most challenges proving intractable without templates. State-of-the-art methods still rely heavily on human expertise and prior knowledge, including structural templates, multiple sequence alignments, and low-resolution experimental data[30,51]. In contrast, RNA-MobiSeq offers a fundamentally different paradigm: it enables fully template/fragment-free modeling by experimentally generating artificial homologous sequences and extracting covariation-derived base-pair restraints without supervision. These restraints can seamlessly interface with both energy-based and machine-learning frameworks, bridging the gap between experimental mutational information and computational RNA folding.

Practically, RNA-MobiSeq is also a rapid, economical, and scalable approach for detecting tertiary interactions and guiding RNA 3D structure prediction. The entire experimental workflow, including EP-PCR library construction, mobility-based screening, and high-throughput sequencing, typically takes about one month, primarily because achieving high double-mutation coverage usually requires three rounds of EP-PCR, and this can be shortened to 2-3 weeks by directly employing doped synthetic libraries, while the downstream computational analysis can be completed within hours. The total cost of gene synthesis, materials, consumables, and sequencing is less than USD 300 per RNA, highlighting RNA-MobiSeq as an accessible experimental–computational framework for RNA

structure inference.

Despite its general robustness, RNA-MobiSeq has several limitations. The experimental workflow may introduce noise during manual gel cutting and generally requires high double-mutation coverage (>80%) for accurate inference, often necessitating multiple rounds of EP-PCR. In addition, current analyses are restricted to single and double mutants and focus on canonical (A-U, G-C) and wobble (G-U) base pairs due to limitation of existing energy functions employed in 2D folding, leaving noncanonical and higher-order interactions underexplored. Finally, intrinsic limitations in the BRiQ energy function and conformational sampling still constrain the precision of 3D structure reconstruction.

## ONLINE CONTENT

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements; details of author contributions and competing interests; and statements of data and code availability are available online.

## REFERENCES


1. Sharp, P. A. The centrality of RNA. *Cell* **136**, 577–580 (2009).
2. Flores, J. K., Walshe, J. L. & Ataide, S. F. RNA and RNA–protein complex crystallography and its challenges. *Aust. J. Chem.* **67**, 1741–1750 (2014).
3. Scott, L. G. & Hennig, M. RNA structure determination by NMR. *Methods Mol. Biol. Clifton NJ* **452**, 29–61 (2008).
4. Ma, H., Jia, X., Zhang, K. & Su, Z. Cryo-EM advances in RNA structure determination. *Signal Transduct. Target. Ther.* **7**, 1–6 (2022).
5. Kretsch, R.C. et al. Naturally ornate RNA-only complexes revealed by cryo-EM. *Nature* (2025). https://doi.org/10.1038/s41586-025-09073-0.
6. Wang, X.-W., Liu, C.-X., Chen, L.-L. & Zhang, Q. C. RNA structure probing uncovers RNA structure-dependent biological functions. *Nat. Chem. Biol.* **17**, 755–766 (2021).
7. Solayman, M. *et al.* Probing RNA structures and functions by solvent accessibility: an overview from experimental and computational perspectives. *Brief. Bioinform.* **23**, bbac112 (2022).
8. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.* **10**, 1643–1669 (2015).
9. Cheng, C. Y. *et al.* Consistent global structures of complex RNA states through multidimensional chemical mapping. *eLife* **4**, e07600 (2015).

10. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* **3**, 954–962 (2011).

11. Das, R., Laederach, A., Pearlman, S. M., Herschlag, D. & Altman, R. B. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA N. Y. N* **11**, 344–354 (2005).

12. Cheng, C. Y., Kladwang, W., Yesselman, J. D. & Das, R. RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc. Natl. Acad. Sci.* **114**, 9876–9881 (2017).

13. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).

14. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).

15. Zhang, Z. *et al.* Accurate inference of the full base-pairing structure of RNA by deep mutational scanning and covariation-induced deviation of activity. *Nucleic Acids Res.* **48**, 1451–1465 (2020).

16. Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14664–14669 (2007).

17. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51–55 (2008).

18. Li, J., Zhang, S., Zhang, D. & Chen, S.-J. Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics* **38**, 4042–4043 (2022).

19. Xiong, Y., Zhang, Y., Wang, J. & Xiao, Y. Using 3dRNA/DNA for RNA and DNA 3D Structure Prediction and Evaluation. *Curr. Protoc.* **3**, e770 (2023).

20. Biesiada, M., Pachulska-Wieczorek, K., Adamiak, R. W. & Purzycka, K. J. RNAComposer and RNA 3D structure prediction for nanotechnology. *Methods San Diego Calif* **103**, 120–127 (2016).

21. Ding, F. *et al.* Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA N. Y. N* **14**, 1164–1173 (2008).

22. Boniecki, M. J. *et al.* SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63 (2016).

23. Xiong, P., Wu, R., Zhan, J. & Zhou, Y. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nat. Commun.* **12**, 2777 (2021).

24. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

25. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

26. Wang, W. *et al.* trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **14**, 7266 (2023).

27. Li, Y. *et al.* Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **14**, 5745 (2023).

28. Shen, T. *et al.* Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods* **21**, 2287–2298 (2024).

29. Das, R. *et al.* Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins* **91**, 1747–1770 (2023).

30. Kretsch, R.C. *et al.* Assessment of nucleic acid structure prediction in CASP16. *bioRxiv* https://doi.org/10.1101/2025.05.06.652459 (2025).

31. Zhou, Y., Litfin, T. & Zhan, J. 3 = 1 + 2: how the divide conquered de novo protein structure prediction and what is next? *Natl. Sci. Rev.* **10**, nwad259 (2023).

32. Schneider, B. *et al.* When will RNA get its AlphaFold moment? *Nucleic Acids Res.* **51**, 9522–9532 (2023).

33. Solem, A. C., Halvorsen, M., Ramos, S. B. V., & Laederach, A. The potential of the riboSNitch in

personalized medicine. *WIREs RNA* **6**, 517-532 (2015).

34. Kutchko, K. M. et al. Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. *RNA* **21**, 1274-1285 (2015).
35. Lai, L. B. *et al.* Structural basis for impaired 5' processing of a mutant tRNA associated with defects in neuronal homeostasis. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2119529119 (2022).
36. Chapman, E. G., Moon, S. L., Wilusz, J. & Kieft, J. S. RNA structures that resist degradation by Xrn1 produce a pathogenic Dengue virus RNA. *eLife* **3**, e01892 (2014).
37. Sun, A. *et al.* SAM-VI riboswitch structure and signature for ligand discrimination. *Nat. Commun.* **10**, 5728 (2019).
38. Stagno, J. R. *et al.* Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. *Nature* **541**, 242–246 (2017).
39. Akiyama, B. M. *et al.* Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science* **354**, 1148–1152 (2016).
40. Gong, B. *et al.* Direct measurement of a pK(a) near neutrality for the catalytic cytosine in the genomic HDV ribozyme using Raman crystallography. *J. Am. Chem. Soc.* **129**, 13335–13342 (2007).
41. Przytula-Mally, A. I. *et al.* Anticodon-like loop-mediated dimerization in the crystal structures of HDV-like CPEB3 ribozymes. 2022.09.22.508989 Preprint at https://doi.org/10.1101/2022.09.22.508989 (2023).
42. Gottipati, K., McNeme, S. C., Tipo, J., White, M. A. & Choi, K. H. Structural basis for cloverleaf RNA-initiated viral genome replication. *Nucleic Acids Res.* **51**, 8850–8863 (2023).
43. Kretsch, R. C. *et al.* Tertiary folds of the SL5 RNA from the 5' proximal region of SARS-CoV-2 and related coronaviruses. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2320493121 (2024).
44. Tipo, J., Gottipati, K., Slaton, M., Gonzalez-Gutierrez, G. & Choi, K.H. Structure of HIV-1 RRE stem-loop II identifies two conformational states of the high-affinity Rev binding site. *Nat. Commun.* **15**, 4198 (2024).
45. Haack, D.B., Rudolfs, B., Jin, S. et al. Scaffold-enabled high-resolution cryo-EM structure determination of RNA. *Nat Commun* **16**, 880 (2025).
46. Zhang, Y. et al. Structural mechanisms for binding and activation of a contact-quenched fluorophore by RhoBAST. *Nat. Commun.* **15**, 4206 (2024).
47. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2017).
48. Weinreb, C. *et al.* 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* **165**, 963–975 (2016).
49. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.* **43**, 10444–10455 (2015).
50. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **6**, 26 (2011).
51. Zhang S, Li J, Zhou Y, Chen SJ. Enhancing RNA 3D structure prediction in CASP16: Integrating physics-based modeling with machine learning for improved predictions. *Proteins.* doi: 10.1002/prot.26856 (2025).

# Figures and Tables

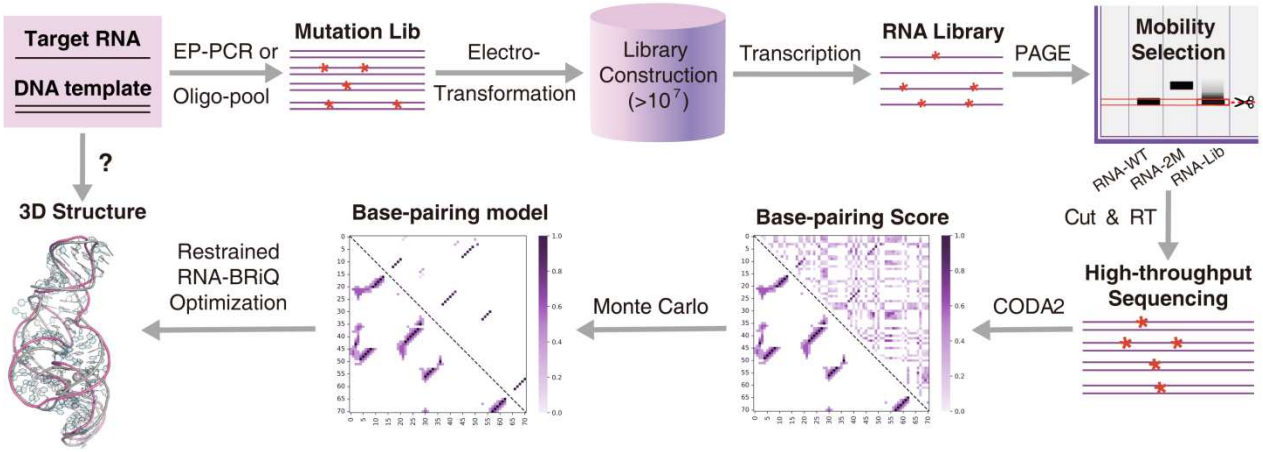



**Figure 1. The workflow of RNA-MobiSeq for RNA 3D structure inference**. The method starts from a DNA template sequence of the target RNA and employs error-prone PCR (EP-PCR) or oligo pools to produce a random-mutation DNA library, followed by *in vitro* transcription to generate an RNA library. The RNA library is then subjected to mobility-based selection by Polyacrylamide Gel Electrophoresis (PAGE). Selected RNAs are reverse transcribed for high-throughput sequencing. Sequencing results are compared to the unselected library for analyzing co-mutation signals by RNA-specific CODA2 (CO-variation-induced Deviation of Activity), following by Monte Carlo (MC) simulated annealing for signal amplification (2D folding). The obtained base pairs were utilized as restraints for predicting RNA structures (3D folding).

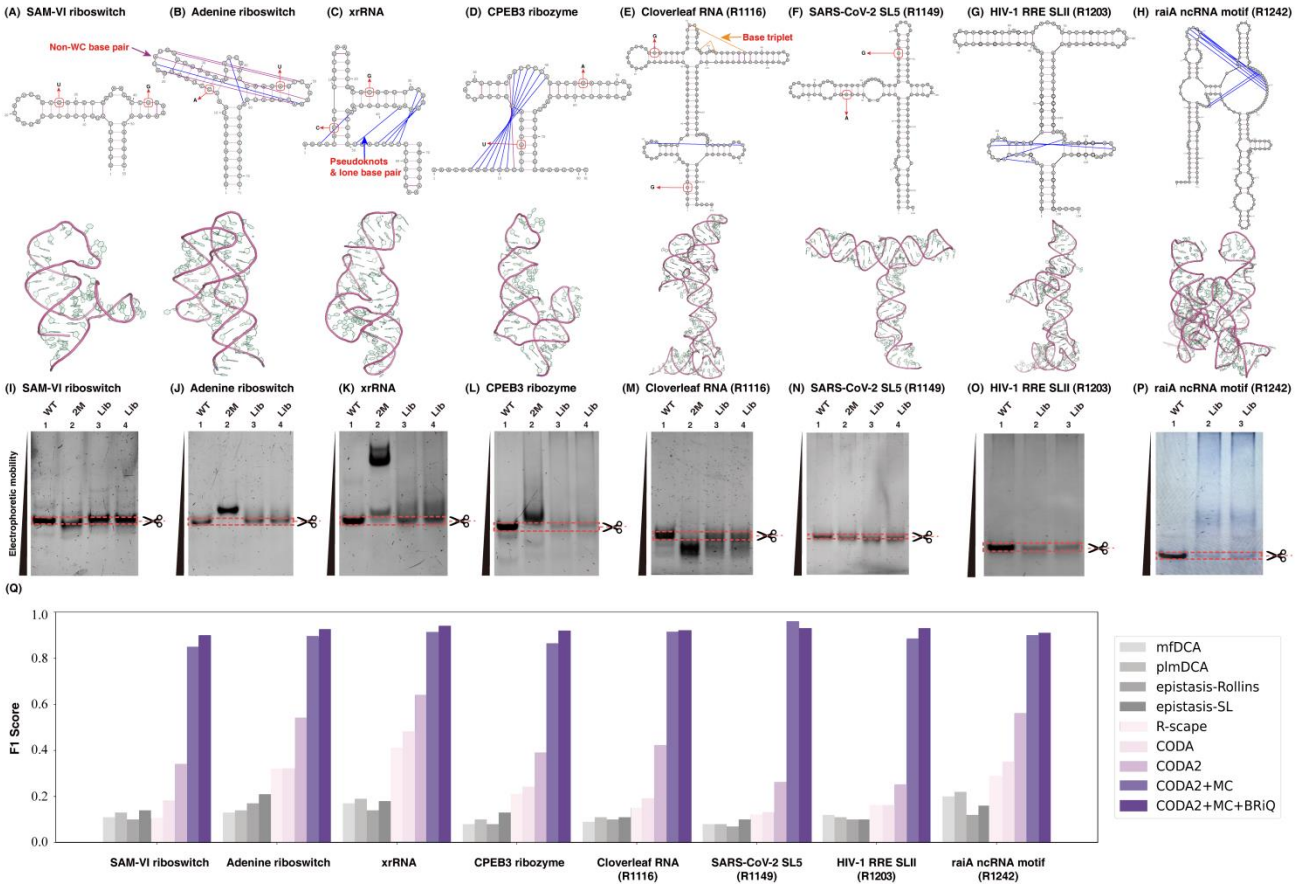



**Figure 2. Mutation-induced changes in mobility and improved extraction of base pairs from mobility-selected sequences by CODA2 and CODA2 + MC over other methods for eight RNAs with different structures.** The secondary and tertiary structures of SAM-VI riboswitch (PDB ID: 6LAS) (A), adenine riboswitch (PDB ID: 5E54) (B), xrRNA (PDB ID: 5TPY) (C), CPEB3 (PDB ID: 7QR4) (D), Cloverleaf RNA (PDB ID: 8S95) (E), and SARS-CoV-2 SL5 (PDB ID: 8UYS) (F), with the locations of randomly chosen double mutations (2M) as indicated (red circles), as well as the secondary and tertiary structures of two CASP 16 targets: RRE SLII (PDB ID: 8UO6) (G), and raiA ncRNA motif (PDB ID: 9ELY) (H). The secondary (mostly local) base pairs are shown in pink, and tertiary (mostly nonlocal) base pairs are shown in blue for pseudoknots and lone base pairs, dark purple for non-Watson-Crick base pairs, and orange for triplets. Mobility comparisons among wild type (WT), double mutants (2M), and two library replicates for SAM-VI riboswitch (I), adenine riboswitch (J), xrRNA (K), CPEB3 ribozyme (L), Cloverleaf RNA (M) and SARS-CoV-2 SL5 (N), as well as mobility comparisons among WT, and two library replicates for two CASP 16 targets: RRE SLII (O) and raiA (P). Bands with the same mobility as the wild-type were excised (highlighted with red dashed boxes). (Q) Comparisons of *F1-scores* inferred by CODA, the RNA-specific CODA (i.e., CODA2), and signal amplification by Monte Carlo simulations (CODA2 + MC, the best in top 5) and further incorporating RNA-BRiQ 3D modeling (CODA2 + MC + BRiQ, the best in top 5), as well as R-scape, the mean-field direct coupling analysis (mfDCA-RNA), pseudolikelihood maximization coupling analysis (plmDCA-RNA), epistasis from Schmiedel & Lehner (epistasis-SL) and from Rollins et al (epistasis-Rollins) for the SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA (R1116), SARS-CoV-2 SL5 (R1149), RRE SLII (R1203), and raiA (R1242), as labelled. *F1-score*=1 (100%) for perfect prediction.

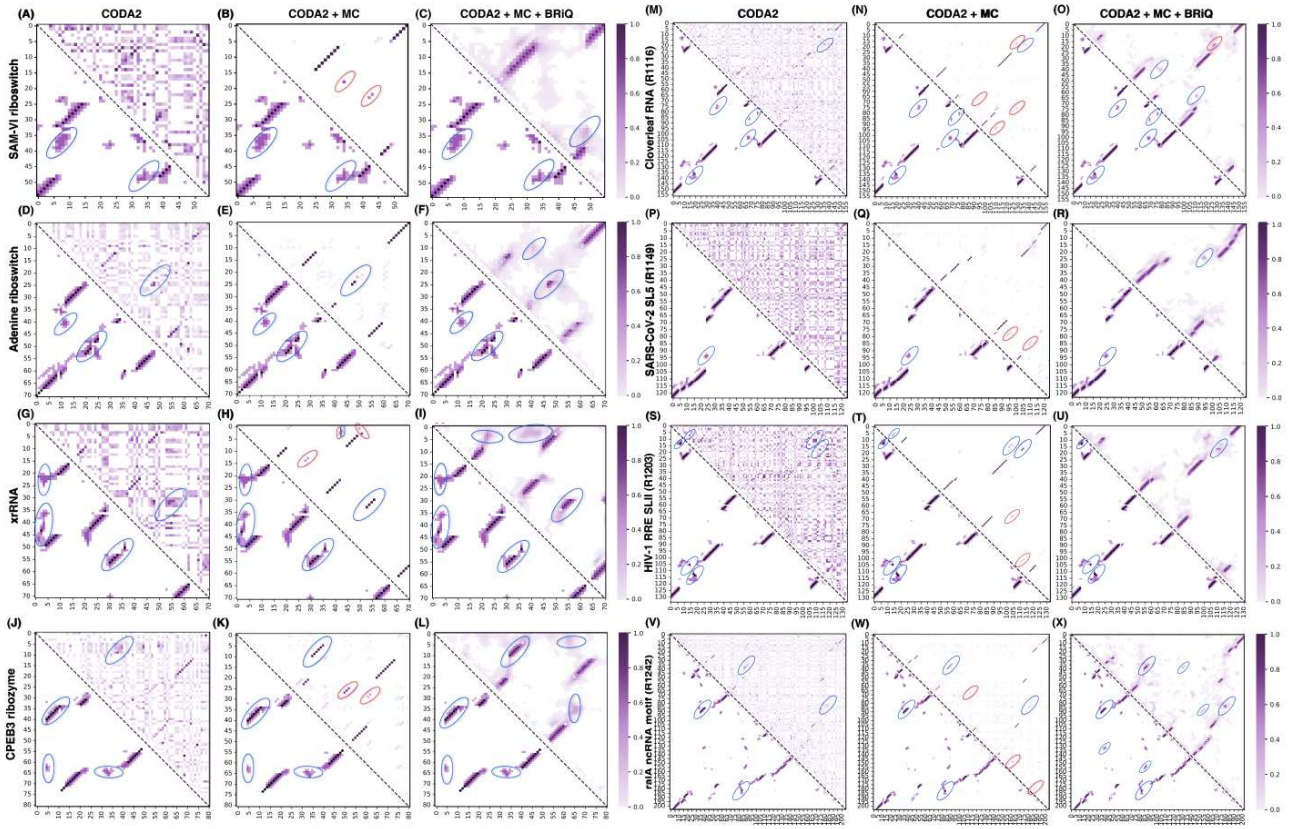



**Figure 3. Detection of tertiary base pairs by CODA2 and subsequent enhancement by CODA2 + MC and CODA2 + MC + BRiQ as indicated by blue circles in upper right triangles.** The base-pair and/or distance-based contact maps (upper right) inferred from CODA2, by MC simulated annealing (CODA2 + MC), and by folding with the BRiQ statistical energy function and the restraints from CODA2 + MC as labelled on the top for SAM-VI riboswitch, adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA ncRNA motif, as labelled on the y-axis. Blue circles: native (lower left triangles) or predicted (upper right triangles) tertiary interactions including pseudoknots, noncanonical base pairs, lone base pairs, and base triplets. Red circles in upper triangles: false positive predictions of base pairs. In the native/predicted distance-based contact maps, the contacts with a 5 Å cutoff are shown in dark purple, while those with an 8 Å cutoff are shown in light purple. The base pairs from the top 100 structures predicted by BRiQ are marked by a gradient of light to dark blue, indicating the probability from low to high.

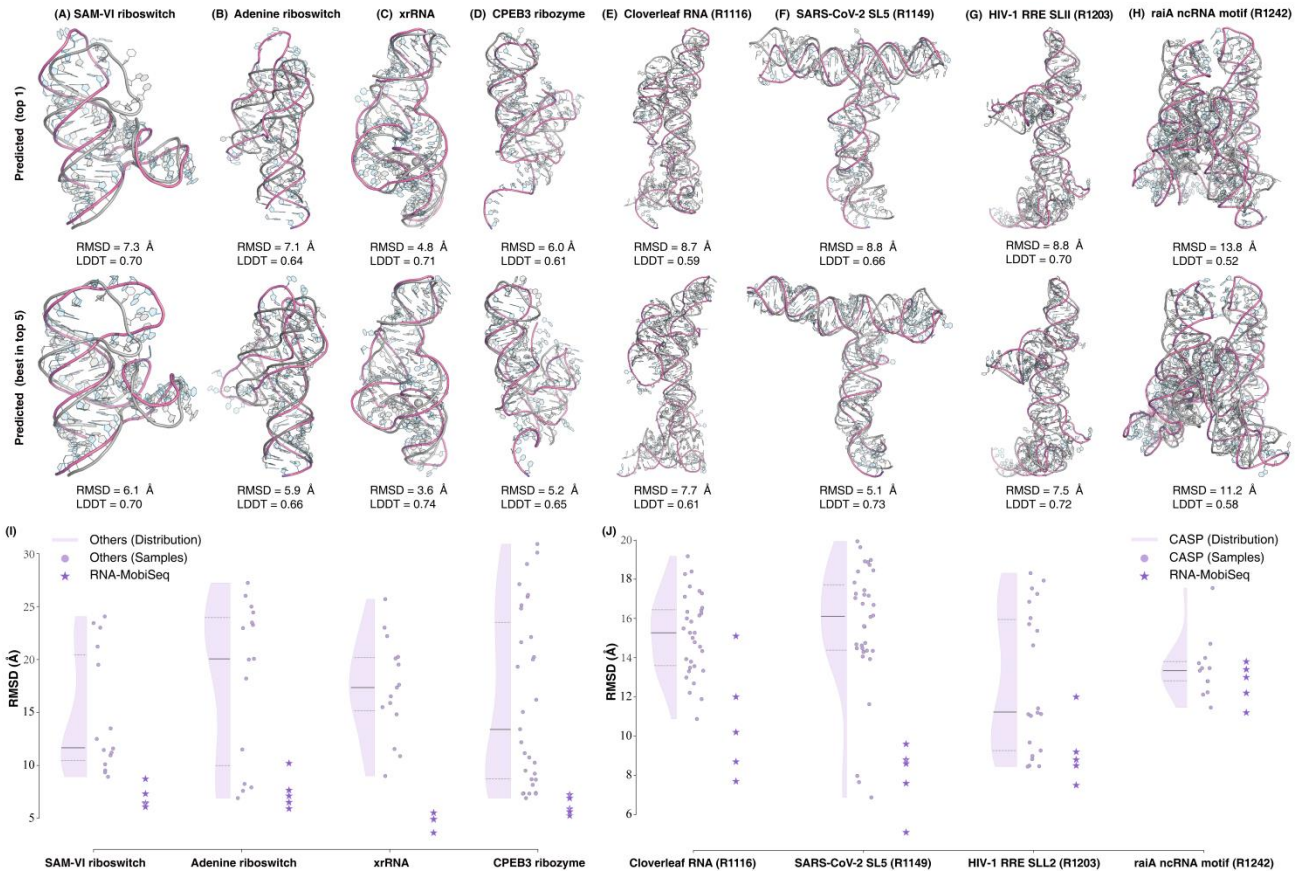
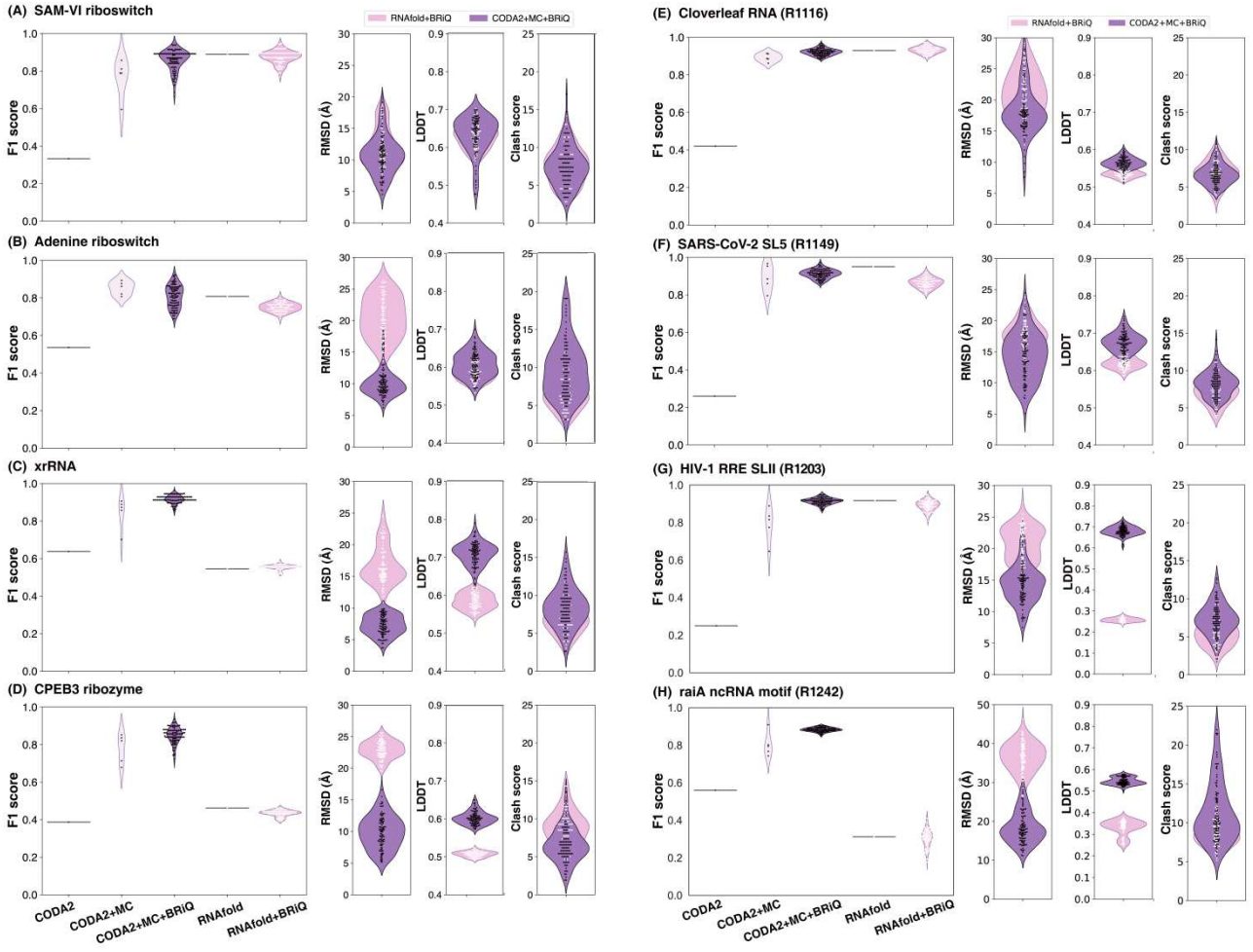



**Figure 4. RNA-MobiSeq accurately infers RNA 3D structures across diverse benchmarks.** (A-H) For all eight RNAs, RNA-MobiSeq achieves ≤ 13.8 Å RMSD and ≥ 0.52 LDDT for top-1 prediction and ≤ 11.2 Å RMSD and ≥ 0.58 LDDT for the best among the top 5 models: (A) SAM-VI riboswitch, (B) adenine riboswitch, (C) xrRNA, (D) CPEB3 ribozyme, (E) Cloverleaf RNA, (F) SARS-CoV-2 SL5 RNA, (G) RRE SLII, and (H) raiA ncRNA motif. Grey: experimental structures. Colored: predicted structures. (I) RMSDs of the top 5 RNA-MobiSeq predictions are compared with those from state-of-the-art RNA structure prediction methods, including energy function-based approaches (Vfold, SimRNA, and MC-Fold/MC-Sym) and deep learning-based approaches (AlphaFold 3, trRosettaRNA, DRfold, and Rhofold), for SAM-VI riboswitch, adenine riboswitch, xrRNA, and CPEB3 ribozyme. (J) RMSDs of the top 5 RNA-MobiSeq predictions are compared with those from independent, template-free methods submitted to CASP15 and CASP16 for four CASP targets: Cloverleaf RNA (R1116), SARS-CoV-2 SL5 RNA (R1149), RRE SLII (R1203), and raiA ncRNA motif (R1242). Stars: the top 5 RNA-MobiSeq predictions. Dots: predictions from other methods or CASP submissions. Violin: the RMSD distributions of the competing methods.




689

690 **Figure 5**. **Folding with base pairs from RNA-MobiSeq can achieve more accurate 3D structures even for**
691 **RNAs having few tertiary base pairs (SAM-VI riboswitch, Cloverleaf RNA, SARS-CoV-2 SL5, and RRE**
692 **SLII)**. *F1-score*s for base pairs inferred from CODA2, CODA2 + MC (Top 5), and CODA2 + MC + BRiQ (Top
693 100) as compared to the reference base pairs predicted by RNAfold and RNAfold + BRiQ as well as the
694 distributions of structure-prediction performance measures for the eight RNAs folded by BRiQ (top 100) based on
695 the secondary structures from CODA2 + MC and RNAfold, respectively, according to RMSD, LDDT, and clash
696 score for SAM-VI riboswitch (A), adenine riboswitch (B), xrRNA (C), CPEB3 ribozyme (D), cloverleaf RNA (E),
697 SARS-CoV-2 SL5 (F), RRE SLII (G), and raiA ncRNA motif (H).

698

699

700

701

702

703 **Table 1**. Comparison of *F1-scores* from MobiSeq with those from other methods that combine
704 experimental techniques with computational algorithms for RNA base-pair inference.

| RNA | Method | *F1-score* | CODA2 | CODA2+MC |
|---|---|---|---|---|
| Adenine riboswitch [a] | Mutate-and-map (SHAPE)[10] | 0.50 | **0.54** | 0.89 |
| xrRNA | M2-Seq (DMS probe)[12] | 0.20 | **0.64** | 0.85 |
| SARS-CoV-2 SL5 (R1149) [b] | M2-Seq (DMS probe + Large library)[43] | 0.14 | **0.26** | 0.96 |
| SARS-CoV-2 SL5 (R1149) [b] | M2-Seq (DMS probe + Scarless)[43] | **0.62** | 0.26 | 0.96 |

[a] The adenine riboswitch (PDB code: 1Y26; sequence: CGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUC UUGAUUAUGAAGUG) used in Ref. 1 and that studied in this work (PDB code: 5E54; sequence: GGGAAGAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACU CUUGAUUAUCUUCCC) differ only by subtle variations in the sequences of 6 nucleotides at each end (marked with underlines), but they have the same length (i.e., 71nt) and similar secondary structure. [b] Results were inferred from 2D chemical mapping (M2-seq) using mutational sequencing-based inference of dimethyl sulfide (DMS) modifications on two different libraries: large-library (using oligonucleotide libraries that encoded SL5 domains as well as all of their single mutants, with 3' barcode hairpins to allow unambiguous deconvolution of the mutant profiles) and scarless (removing primer binding sequences from the RNA to prevent unwanted secondary structure interference during DMS modification)

705

706

707

708

709

# METHODS

## DNA Templates, RNA Synthesis, and Purification

The corresponding DNA templates of the eleven RNAs were directly synthesized and subcloned into a vector named as pUC57-T7Q-RNA (pUC57-1.8k vector with N-terminal T7 RNA polymerase promoter/ adaptor sequence-1 and C-terminal adaptor sequence-2/BspQ I restriction enzyme site). The detailed sequence information is listed in Supplementary Table S2.

The restriction enzyme BspQI was utilized to linearize the DNA-template plasmids. The linearized plasmids were agarose-gel purified and then employed as the templates for *in vitro* RNA synthesis. Each reaction system contains 1× RNA Pol reaction buffer (NEB, B9012), 5 U/μL T7 RNA polymerase (NEB, M0251), 1 U/μL RNase inhibitor (NEB, M0314), 1 mM NTPs (NEB, N0416) and 40 ng/μL DNA template. After transcription reaction at 37 °C for 5 h, DNase I (NEB, M0303) was added to the reaction system to a final concentration of 0.05 U/μL. The reaction system was further incubated at 37 °C for 30 minutes to eliminate the DNA template. The synthesized RNAs were purified using a spin-column kit (RNA Clean & Concentrator-5, Zymo Research, R1016) according to the manufacturer's instructions. The concentrations of RNAs were determined using NanoDrop One (Thermo Scientific). The purity of all newly synthesized RNAs was detected using Urea-PAGE which was stained with SYBR Gold (Invitrogen, S11494). The gels were finally visualized on a ChemiDoc MP imaging system (Bio-Rad).

## Preparation of DNA Mutant Libraries

The wild type DNA sequences of the selected RNAs were mutated by EP-PCR using QuickMutation Random Mutagenesis Kit (Beyotime, D0219) according to the manufacturer's instructions with primers M13-FP and M13-RP (Supplementary Table S5). A total of three rounds of EP-PCR were performed to reach the desired mutation rate. The products from the previous EP-PCR round were purified using Urea-PAGE and used as the templates for the next EP-PCR round. The products from

the last EP-PCR round were purified as the EP-PCR DNA mutant libraries. The target DNA bands with expected sizes were excised, crushed and frozen at -80 °C. The frozen solidified products were quickly thawed in a 90 °C hot bath for 5 mins and the target DNA was passive eluted in TE/NaCl buffer (10 mM Tris–HCl, 1mM EDTA, pH 8.0, 100 mM NaCl) on a rotary shaker at 37 °C for overnight. The DNAs were finally precipitated using ethanol and redissolved in nuclease-free water.

In addition to EP-PCR, all single and double mutants of the two RNA targets R1116 and R1149 of CASP15 were also synthesized as oligo pools (Genscript). This method reduces the need for tedious experimental steps (EP-PCR reactions and Urea-PAGE purifications) and saves time. The synthesized oligo pools were also amplified using M13-FP and M13-RP and purified as the oligo-pool DNA mutant libraries.

The EP-PCR and oligo-pool DNA mutant libraries were amplified with primers M13-Plus-FP/M13-Plus-RP and the linearized pUC57-T7Q vector was amplified with primers AS-FP/AS-RP using PrimerSTAR HS DNA polymerase (Takara, R040A) (Supplementary Table S5). The DNA fragments were agarose-gel purified and then assembled with the linearized pUC57-T7Q vector using CloneExpress II One Step Cloning Kit (Vazyme, C112-02). The assembled products were purified with the RNA Clean & Concentrator-5 (Zymo Research, R1016) and eluted with nuclease-free water.

The assembled products were transformed into electrocompetent DH5α cells using an MicroPulser electroporation instrument (Bio-Rad). The transformants were cultured on 20 mL LB for 1 h at 37 °C on a rotator with 220 rpm. After 1 h, the size of the constructed library was detected by the number of colonies after a gradient dilution on LB-agar plate (containing 100 μg/mL ampicillin). Then 100 μg/mL ampicillin was added to 20 mL LB and allowed to grow at 37 °C for 10 h and collected using centrifugation. The pellet was resuspended in LB medium containing 15% glycerol and stored at -80 °C for long term use. Ten clones were randomly selected and sent for Sanger sequencing to

evaluate the percentage of clones with the right insertion form in our libraries. The size capacities of constructed libraries were all greater than $1 \times 10^7$.

**Mobility-based Selection**

$10^7$ clones from the mutated library were cultured and amplified. The corresponding plasmids were extracted. The BspQI linearized vectors were transcribed to RNAs. RNA samples were heated to 70 °C for 5 min and slowly cooled to 4 °C in the folding buffer containing 10 mM $MgCl_2$, 10 mM KCl, and 10 mM sodium acetate (pH 6.8). RNA samples were mixed with nondenaturing loading buffer (TBE buffer, 4% glycerol, 0.01% bromophenol blue, 0.01% Xylene cyanol) and were finally loaded to 7.5% or 10% native page gels (the gel concentration chosen according to the size of the RNA) in the TBE buffer for running at 60 V for 10 - 12 h at ice bath equipment in a cold room. The bands, which showed the same mobility as wild type RNA sequences, were excised and RNAs were extracted and purified using the method as DNA extraction in PAGE gel, as described above.

**Sequencing Library Preparation and High-throughput Sequencing (HTS)**

Each purified RNA band was reverse transcribed with the following reaction system: 1 × ProtoScript II buffer, 10 mM DTT, 0.5 mM dNTPs, 0.4 U/μL RNase inhibitor (NEB, M0314), 10 U/μL ProtoScript II reverse transcriptase (NEB, M0368), 2 μL M13-RP and the RNA template. The reaction system was incubated at 42 °C for 1 h and then inactivated at 65 °C for 20 min. The cDNA product was stored at -20 °C for further use. A 10 μL cDNA product was utilized as the template for amplification (10 cycles) with primers M13-FP/M13-RP. Then, the DNA sequence library was constructed using recycled fragments as the template with adaptor primers (Supplementary Table S5) in PCR amplification reactions. The PCR products were agarose-gel purified. The DNA concentration was measured using an Invitrogen Qubit Fluorometer and the peak size was analyzed by an Agilent 2100 Bioanalyzer. The DNA samples were high throughput sequenced using Illumina NovaSeq in PE150 mode at Novogene Technology Co., Ltd.

## HTS Data Analysis

HTS data analysis followed a procedure similar to our previous work on self-cleaving ribozymes[15,52]. Paired-end sequencing reads were first merged into single reads by using PEAR[53]. Then, we counted the number of reads for all the mutants and the wild type before and after mobility-based selection. The fitness $f_i$ of a mutant $i$ is calculated based on its read counts before ($n_i{}^{pre}$) and after ($n_i{}^{post}$) mobility-based selection, normalized by the read counts of the wild type ($n_{wt}{}^{pre}$ and $n_{wt}{}^{post}$, respectively), *i.e.*,

$$f_i = \frac{n_i{}^{post}/n_{wt}{}^{post}}{n_i{}^{pre}/n_{wt}{}^{pre}}, \quad (1)$$

## RNA-specific CODA Analysis for Mobility-selected RNAs

CODA provides unsupervised searches for covariation-induced deviation (outliers) of the fitness score of a double mutant from the independent mutation model when two mutations were treated as independent from each other[15]. To establish an independent-mutation model, we built a dataset for all single and double mutants, with the corresponding fitness-scores from high-throughput sequencing. Unlike the original CODA model, which employed the fitness scores of two single mutants as the only input features, the CODA2 incorporated six additional features for each double mutant: mutational (*mt*) base types and the sequence positions of these two mutations (*pos*), and wild (*wt*) to better describe the mutation characteristics. Base types were encoded with one-hot encoding (1000, 0100, 0010, and 0001 for A, C, G, and U, respectively). After max-min normalization of the data, the model was constructed using a support vector machine regression (SVR) to fit the fitness scores of all double mutants as a function of the features of their corresponding single mutants, that is, $f(X,Y) = G(pos(X), pos(Y), wt(X), wt(Y), mt(X), mt(Y), f(X), f(Y))$, where $X$ and $Y$ are the two corresponding single-mutants in the double mutant. Then, the fitness score for any double mutants can be predicted by the trained SVR model, and the CODA score was calculated through the

deviations between predicted ($f_{pred}$) and actual ($f_{act}$) fitness scores of the double-mutant, $\mathrm{CODA} = (f_{act} - f_{pred})/(f_{pred} + 0.2)$, where the shift of 0.2 is employed to avoid near zero value of $f_{pred}$. This regression equation is different from the original CODA by including mutation positions and base types before and after mutations as the features.

As in CODA, the pairing score (*PS*) of a double mutation was calculated by a two-state naive Bayesian classifier with

$$PS = \log \frac{p(CODA|paired)}{p(CODA)}, \quad (2)$$

where $p(CODA) = p(paird) \times p(CODA|paired) + (1 - p(paird)) \times p(CODA|unpaired)$, and the probability is derived from a mixture model of two Gaussian distributions without using a training set[15]. For the mixture model, the first Gaussian distribution of CODA score was fitted from all double mutation data. Then, the second distribution was obtained by the subset of data points with CODA score $> \mu_1 + n\sigma_1$, based on the expectation ($\mu_1$) and standard deviation ($\sigma_1$) of the first distribution, where *n* is an adjustable parameter; see Fig. S2 and Supplementary Fig. S1.

The parameters (i.e., C and γ) in SVR, as well as the above adjustable parameter n (i.e., the variance truncation coefficient), generally depend on the distribution of different RNA mutation data. Thus, CODA2 performed a grid search on the parameters for each dataset, and calculated the energy of inferred base pairs under different parameter settings using the thermodynamic parameters[54] (see Eq. 3 in the next section). The contacts with the lowest free energy are then selected as the final inference result; see Fig. S2. Compared to the fixed parameters used in CODA, these RNA-specific parameterizations in CODA2 enhanced the model's robustness for RNAs with different lengths and having sequence-dependent distributions of mutations.

**Base-pair Prediction by MC Simulated Annealing (CODA2+MC)**

As in our previous work[15], we further enhanced pairing signals by integrating the above pairing score

826 *PS* with experimental free energies for stacking by

$$E = PS + \omega * \left(\Delta G_{stacking} + \Delta G_{AU_end_penalty} + \Delta G_{lone}\right), \quad (3)$$

827 where $\omega$ is an optimized weight (0.9 in this work). $\Delta G_{stacking}$ in Eq. (3) is the sequence-dependent
828 stacking energy between two adjacent base pairs, calculated by $\Delta G_{stacking} = \Delta H_{stacking} -$
829 $T\Delta S_{stacking}$, where $\Delta H_{stacking}$ and $\Delta S_{stacking}$ are from the thermodynamic measurements, and $T$ is
830 real temperature in Kelvin. $\Delta G_{AU_end_penalty}$ is the penalty of each helical AU end, and $\Delta G_{lone}$
831 (0.5kcal/mol) is the penalty for a lone base pair[54,55]. Unlike the previous work[15], we did not add a
832 penalty for noncanonical base pairs but instead ignored them in MC simulated annealing for a
833 smoother energy landscape.

834 In MC simulated annealing, the initial temperature was set at 120℃ and decreased by a factor of
835 0.95 until temperature reached ~25℃. At each temperature, $1\times10^5$ MC steps were performed. In each
836 MC step, a random new base pair (*i*-*j*) between bases *i* and *j* (|*j*-*i*|>3) was generated while the existing
837 pairs related to the new pair were destroyed, and the change was accepted or rejected according to
838 the Metropolis criterion[55,56]. At the end of the simulated annealing procedure, the secondary structure
839 with the lowest energy was regarded as the prediction. This annealing procedure could run
840 independently multiple times (e.g., 1000), and after clustering, the top 5 base-pair structures were
841 taken as the final predicted results.

842 **Tertiary Structure Prediction Using the RNA-BRiQ Statistical Potential**

843 We further employed the RNA-BRiQ statistical energy function[23,57] for building 3D structures for
844 RNAs. The base-pair structures predicted above (CODA2 + MC) from deep mutational scanning data
845 as well as the sequence of an RNA were inputted for energy minimization by a nucleobase-centric tree
846 algorithm with default parameters. For each of the top 5 predicted base-pairing configurations, 1000
847 3D models were generated using the RNA-BRiQ with different random seeds. For RNAs longer than

120nt, five rounds of sampling were performed. Then, all generated models were further scored by the BRiQ energy, and the 100 models with the lowest energy scores (i.e., top 100) were taken as the final predicted tertiary structures.

**Performance Evaluation**

Predicted base pairs were evaluated by *F1 score* (*F*1), which is the harmonic mean of sensitivity (*Sen*) and precision (*Pre*) as below:

$$F1 = \frac{2 \times Sen \times Pre}{Sen + Pre}, \tag{4}$$

where $Sen = TP/(TP + FN)$ and $Pre = TP/(TP + FP)$. Here, *TP, FN, FP,* and *TN* are the numbers of true positive, false negative, false positive, and true negative contacts in prediction, respectively. A balanced measure is the Matthews Correlation Coefficient (*MCC*) calculated as follows[58]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{5}$$

Since the probability of any two bases pairing is much smaller than the probability of them not pairing, i.e., the number of negative samples far exceeds that of positive samples, resulting in *TN* typically being much larger than *TP, FN*, and *FP*, the MCC can be approximated as $\text{MCC} \approx \frac{TP \times TN}{\sqrt{(TP+FP)(TP+FN) \times TN \times TN}} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}$, which is similar to the *F1-score* as we found here. The base pairs of native structures and structures predicted by BRiQ were derived by the DSSR software[59], and the maximal *F1-score* was used to characterize the accuracy of predicted contacts.

As in CASP 15, we evaluated 3D predicted structures by RMSD (root mean square deviation) and LDDT (local distance difference test)[29,30,60,61]. The LDDT (unlike for proteins, the implementation of LDDT for RNA did not penalize for stereochemical violations) was calculated by OpenStructure software (http://www.openstructure.org)[62]. In addition, we employed clash score for assessing the atomic harmony of the structure[29,63]. The clash score was calculated by MolProbity

869 (http://molprobity.biochem.duke.edu/)[64].

870 **Mutational coupling analysis and epistasis**

871 R-scape (http://eddylab.org/R-scape/)[47] was used to identify evidence of conserved RNA structures
872 by measuring pairwise covariations observed in MSAs generated from RNA double mutants. For
873 each RNA, six MSAs were first generated based on the sequences (reads count > 5 and fitness >
874 cutoff) with double mutants, using cutoff values ranging from 0.5 to 1.5 at intervals of 0.2. Then, for
875 each MSA, R-scape was iterated over different E-values (0.05, 0.1, 1.0, 10.0, 100.0, 1000.0). Finally,
876 based on the returned pairing data, the *F1-score* was calculated for each parameter and the best result
877 was reported.

878 Functional mutants with high fitness scores were used as the input sequences for covariation analysis
879 by mfDCA-RNA and plmDCA-RNA; see more details in Ref[15]. The pydca, a standalone
880 Python-based software package for the DCA of protein- and RNA-homologous families
881 (https://github.com/KIT-MBS/pydca)[65] was used to infer coevolutionary couplings between pairs of
882 nucleotides, based on two popular inverse statistical approaches, namely, the mean-field
883 (mfDCA-RNA) and the pseudo-likelihood maximization (plmDCA-RNA) with default parameters.

884 The epistasis was calculated from a non-parametric null model following Schmiedel & Lehner
885 (epistasis-SL)[13] and Rollins et al. (epistasis-Rollins)[14]. In epistasis-SL, the combined score from the
886 three interaction scores was used to estimate the epistatic interactions
887 (https://github.com/jschmiedel/DMS2structure).

888 **3D structure predictions from other methods**

889 RNA-MobiSeq was compared with representative state-of-the-art RNA 3D structure prediction
890 methods, including four deep learning–based approaches (AlphaFold 3, trRosettaRNA, DRfold, and
891 Rhofold) and three energy function–based approaches (Vfold, SimRNA, and MC-Fold/MC-Sym).
892 For all methods, RNA sequences were used as the sole input, without providing additional structural

constraints unless explicitly required by the method. For AlphaFold 3 (https://alphafoldserver.com) and DRfold (https://zhanggroup.org/DRfold/), predictions were generated using the respective web servers with default settings. Rhofold (https://github.com/Dharmogata/RhoFold) and trRosettaRNA (https://yanglab.qd.sdu.edu.cn/trRosettaRNA/; v1.0), were run using the pretrained standalone models in the single-sequence setting. For energy function–based methods, Vfold predictions were generated using the Vfold pipeline (http://rna.physics.missouri.edu/vfoldPipeline/) with default parameters, enabling H-type pseudoknot consideration and allowing up to five candidate secondary structures for 3D prediction. SimRNA predictions were obtained using SimRNAweb 2.0 (https://genesilico.pl/SimRNAweb/) with default parameters, and the representative structures of the five returned clusters were used as the top five predictions. For MC-Fold/MC-Sym, secondary structures were first predicted using MC-Fold via the MC-Pipeline web service (https://major.iric.ca/MC-Pipeline/) with the option "Consider H-type pseudoknots" enabled. The resulting candidate secondary structures were converted into 3D models using MC-Sym, and the top five models were selected based on the built-in MC-Sym scoring function.

The CASP prediction results for four CASP RNA targets (Cloverleaf RNA (R1116) and SARS-CoV-2 SL5 RNA (R1149) from CASP15, and RRE SLII (R1203) and raiA (R1242) from CASP16) were obtained directly from the official CASP repositories (https://predictioncenter.org/). To ensure a fair and methodologically consistent comparison with RNA-MobiSeq, which does not rely on structural templates, we curated the CASP submissions based on method descriptions provided in the CASP15 and CASP16 abstract books (CASP15: https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf; CASP16: https://predictioncenter.org/casp16/doc/CASP16_Abstracts.pdf). Specifically, submissions that explicitly used structural templates or template-based homology modeling strategies were excluded. In addition, non-independent or ensemble-based approaches were removed, such as PerezLab_Gators, which integrated predicted structures from multiple existing methods (e.g., RNAComposer,

trRosettaRNA, and AlphaFold3). Accordingly, the following groups were excluded from CASP15: AIchemy_RNA2, Chen, CoMMiT-human, CoMMiT-server, FoldEver, FoldEver-Hybrid, GWxraylab, Kiharalab, Kiharalab_Server, LCBio, PerezLab_Gators, RNApolis, Rookie, SHT, SoutheRNA, UltraFold, UltraFold_Server, Yang, and Yang-Multimer; and from CASP16: BRiQX, CSSB_experimental, falcon2, GeneSilico, GuangzhouRNA_AI, GuangzhouRNA-human, GuangzhouRNA-meta, GromihaLab, isyslab-hust, KiharaLab, kiharalab_server, LCBio, MIEnsembles-Server, NKRNA-s, PerezLab_Gators, RNA_Dojo, RNAFOLDX, RNApolis, SoutheRNA, UNRES, Vfold, and Zheng. Furthermore, submissions that were not described in the official CASP abstract books were excluded, as it was not possible to verify whether these methods were independent and template-free. The excluded groups included nucE2E in CASP15, and 405, 406, AIR, B-LAB, dfr, Diff, Dokholyan, FrederickFolding, mmagnus, nfRNA, OmniFold, ShanghaiTech-human, and ShanghaiTech-server in CASP16.

## Data Availability

Illumina sequencing data generated in this study are deposited at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE276399. The top 5 predicted 3D structures of the RNA-Puzzles targets PZ51 and PZ62 generated by RNA-MobiSeq are available at Zenodo (https://zenodo.org/records/15320479 and https://zenodo.org/records/16750052).

## Code availability

The source codes for RNA-MobiSeq are available at GitHub: https://github.com/RNA-folding-lab/RNA-MobiSeq.

## References


52. Zhang, Z. *et al.* Minimal twister sister-like self-cleaving ribozymes in the human genome revealed by deep mutational scanning. *eLife* **12**, RP90254 (2024).
53. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinforma. Oxf. Engl.* **30**, 614–620 (2014).

54. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).

55. Shi, Y.-Z., Jin, L., Feng, C.-J., Tan, Y.-L. & Tan, Z.-J. Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions. *PLOS Computat. Biol.***14**, e1006222 (2018).

56. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).

57. Chen, K., Zhou, Y., Wang, S. & Xiong, P. RNA tertiary structure modeling with BRiQ potential in CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1771–1778 (2023).

58. Parisien M, Cruz JA, Westhof E, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*. **15**:1875-85 (2009).

59. Lu, X.-J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142 (2015).

60. Gong, S., Zhang, C. & Zhang, Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinforma. Oxf. Engl.* **35**, 4459–4461 (2019).

61. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinforma. Oxf. Engl.* **29**, 2722–2728 (2013).

62. Biasini, M. *et al.* OpenStructure: a flexible software framework for computational structural biology. *Bioinforma. Oxf. Engl.* **26**, 2626–2628 (2010).

63. Magnus, M. *et al.* RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res.* **48**, 576–588 (2020).

64. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.* **27**, 293–315 (2018).

65. Zerihun, M. B., Pucci, F., Peter, E. K. & Schug, A. pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics* **36**, 2264–2265 (2020).

## Acknowledgements

This research received support from the National Natural Science Foundation of China (Grant #22350710182) and National Key Research and Development Program of China (NO.2021YFF1200400). It was also benefited from access to the supercomputing resources at Shenzhen Bay Laboratory.

## Author contributions

JZ and JT conducted experimental design. JT performed wet-lab experiments. YS performed data analysis, computational prediction and method development. ZZ and DL participated initial data analysis and method development. JZ and YZ initiated and supervised the project. YZ provided the funding support. JT, YS, and YZ drafted the whole manuscript. All authors involved subsequent manuscript improvement.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper online.

**Supplementary information**. The online version contains supplementary material available at xxx.

**Correspondence and requests for materials** should be addressed to Yaoqi Zhou or Jian Zhan.

# Supplemental Figures

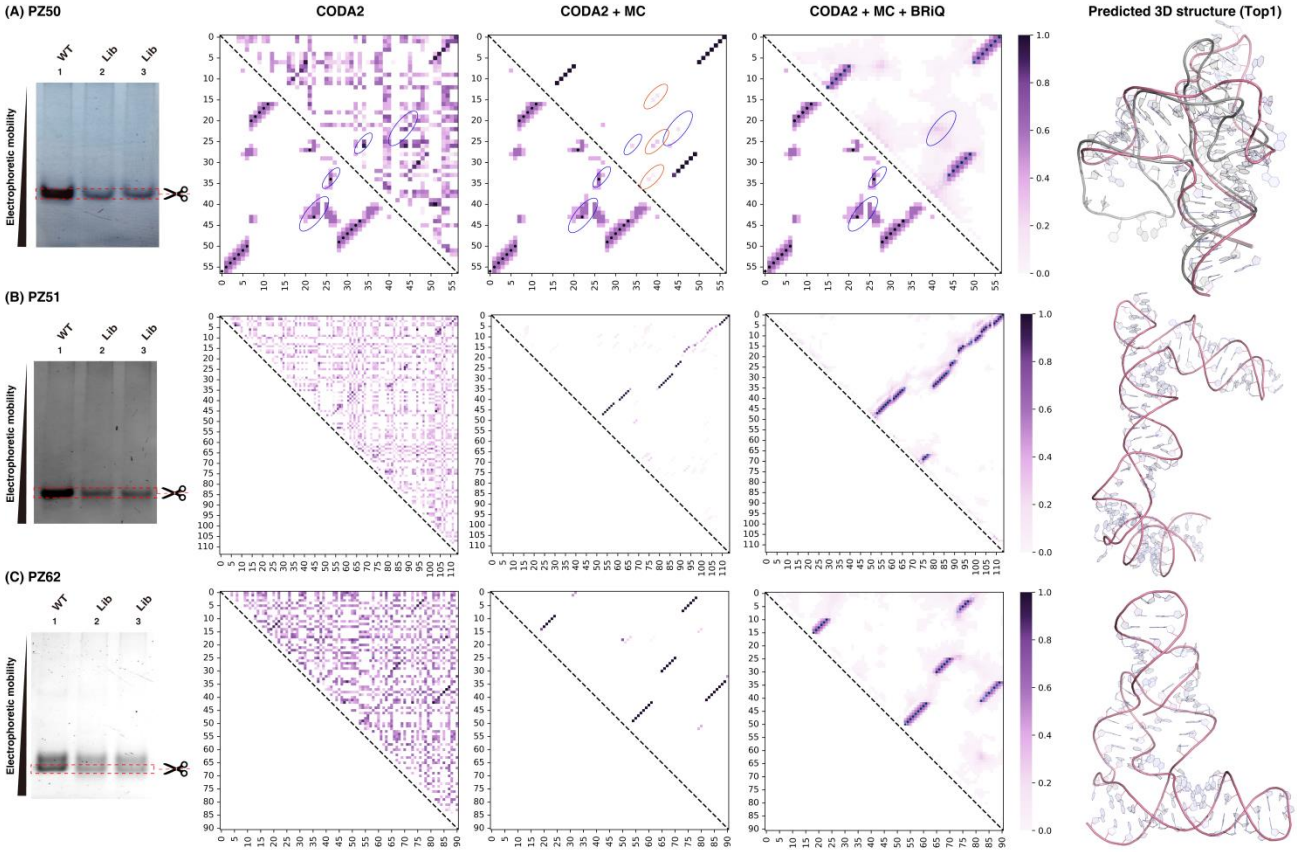



**Figure S1. Mutation induced changes in mobility, and base-pairing as well as 3D structures inferred by RNA-MobiSeq for RNA-Puzzles targets: PZ50 (A), PZ51 (B), and PZ62 (C).** Left: Mobility comparisons among wild type (WT) and two library replicates for the three targets. Middle: The base pair and/or distance-based contact maps (upper right) inferred from CODA2, by MC simulated annealing (CODA2 + MC), and by folding with the BRiQ statistical energy function and the restraints from CODA2 + MC as labelled on the top for the three targets. For PZ50 in (A), the native distance-based contact maps (lower left triangles) are also shown. In the native/predicted distance-based contact maps, the contacts with a 5 Å cutoff are shown in dark purple, while those with an 8 Å cutoff are shown in light purple. The base pairs from the top 100 structures predicted by BRiQ are marked by a gradient of light to dark blue, indicating the probability from low to high. The *F1-scores* inferred by CODA2, CODA2 + MC, and CODA2 + MC + BRiQ are 0.60, 0.87/0.89 (top1/best in the top 5), and 0.90/0.91 (top1/best in the top 5), respectively. Right: The top 1 3D structures predicted by RNA-MobiSeq for the three targets. Grey: experimental structures. Colored: predicted structures.

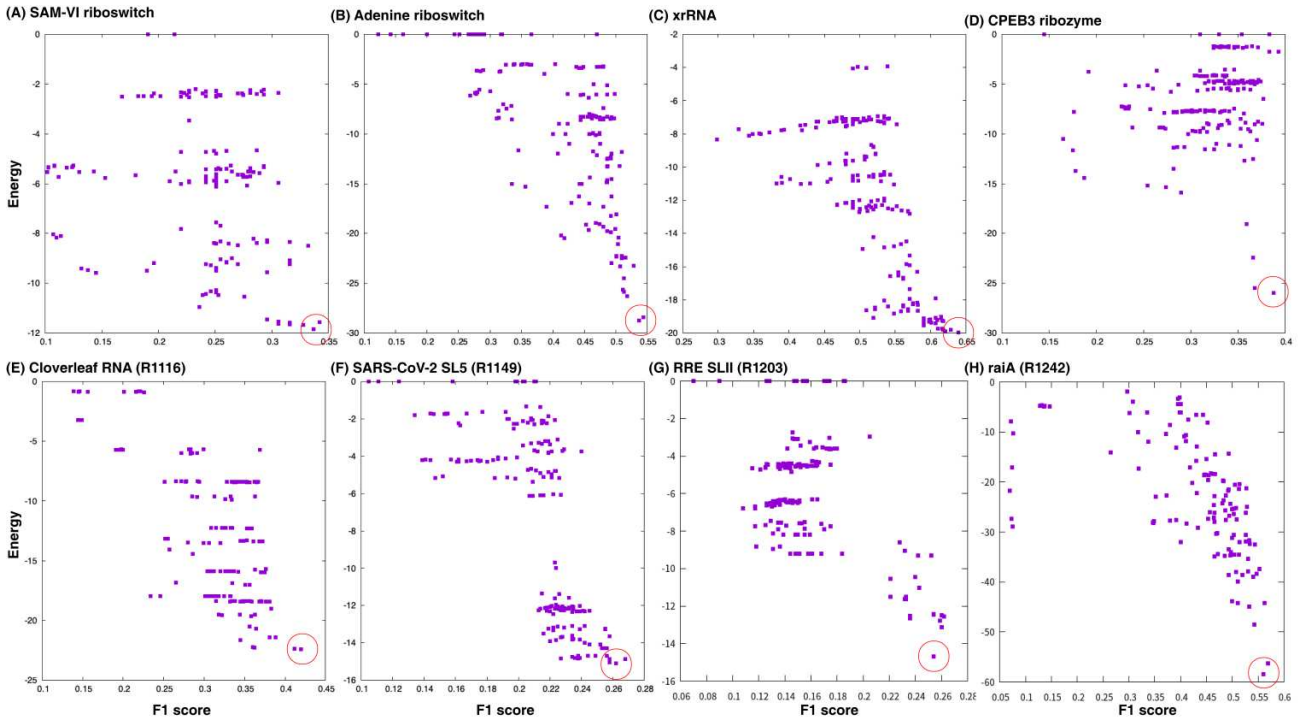



**Figure S2. RNA-specific CODA2 parameters.** Unlike the original version of CODA, to increase the adaptability of the model to different data distributions, we iterated through several adjustable parameters (i.e., C and γ in SVR model, and the variance truncation coefficient *n*) to obtain a series of contact maps. These maps were further scored by the energy calculated by Eq. 3 in the main text using the experimental thermodynamic parameters[5], and the one with the lowest energy was taken as the prediction. (A-H) Energy versus *F1* score of the maps inferred by CODA2 with different parameters for six RNAs: SAM-VI riboswitch (A), Adenine riboswitch (B), xrRNA (C), CPEB3 ribozyme (D), Cloverleaf RNA (E), SARS-CoV-2 SL5 RNA (F), RRE SLII (G), and raiA ncRNA motif (H). The optimized $(C, \gamma, n)$ parameters for the results with the lowest energy are (1.0, 1.0, 2.5), (0.01, 0.1, 3.5), (5.0, 10.0, 3.0), (100.0, 1.0, 4.0), (1.0, 10.0, 3.0), (10.0, 1.0, 1.5), (1.0, 10.0, 4.0), and (1000.0, 0.01, 3.5) for SAM-VI riboswitch, Adenine riboswitch, xrRNA, CPEB3 ribozyme, Cloverleaf RNA, SARS-CoV-2 SL5 RNA, RRE SLII, and raiA ncRNA motif, respectively.

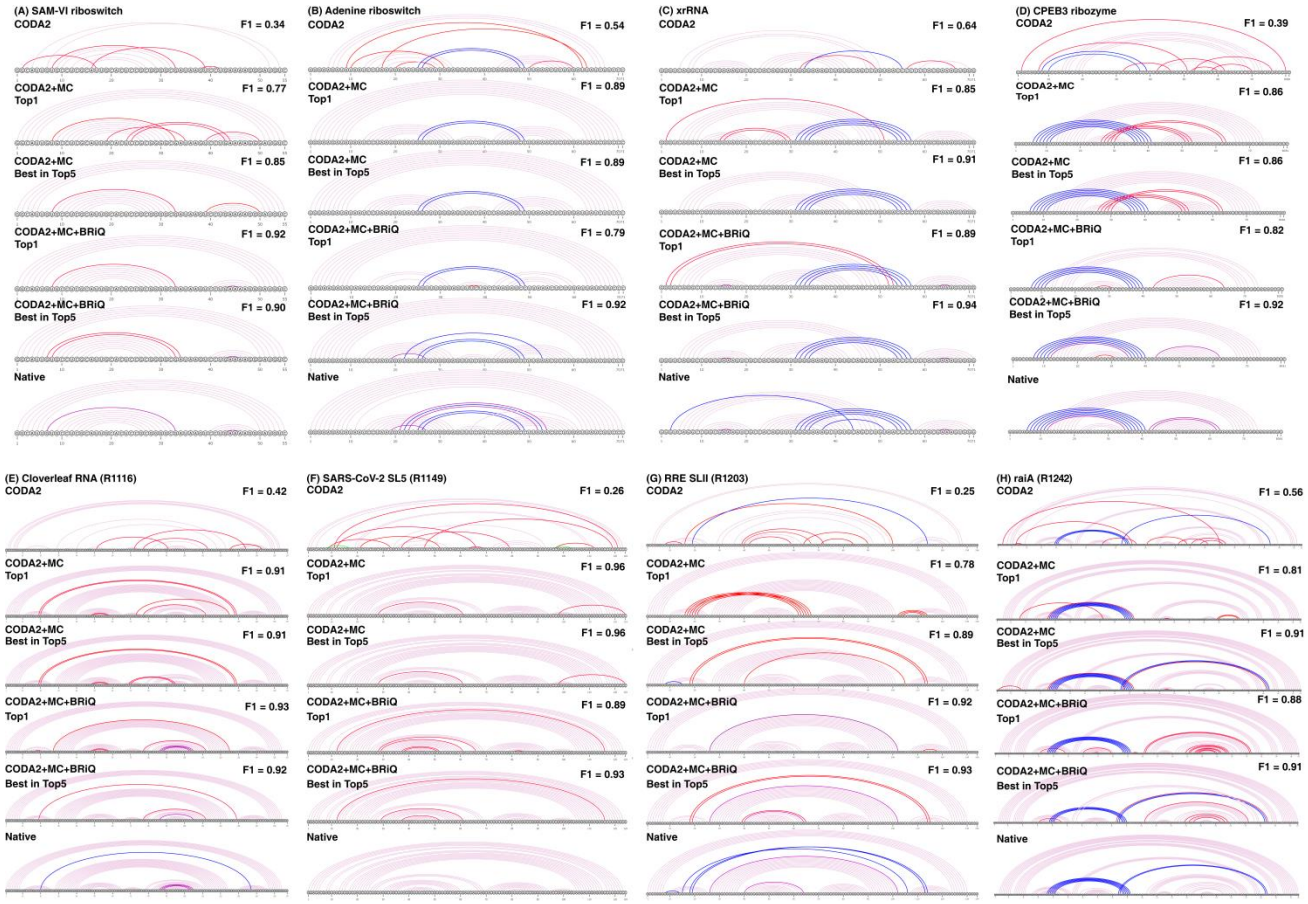



**Figure S3**. **Capture of tertiary base pairs by RNA-MobiSeq**. The comparisons between base pairs derived from CODA2 (top 1), CODA2 + MC (top 1), CODA2 + MC (best in top 5), CODA2 + MC + BRiQ (top 1), CODA2 + MC + BRiQ (best in top 5), and native structures for SAM-VI riboswitch (A), adenine riboswitch (B), xrRNA (C), CPEB3 ribozyme (D), Cloverleaf RNA (E), SARS-CoV-2 SL5 (F), RRE SLII (G), and raiA (H). Secondary (mostly local) base pairs are shown in light purple, whereas tertiary (mostly nonlocal) base pairs such as pseudoknots and lone base pairs are shown in blue, non-Watson-Crick base pairs in dark purple, and base triplets in orange. False positive predictions are shown in red.

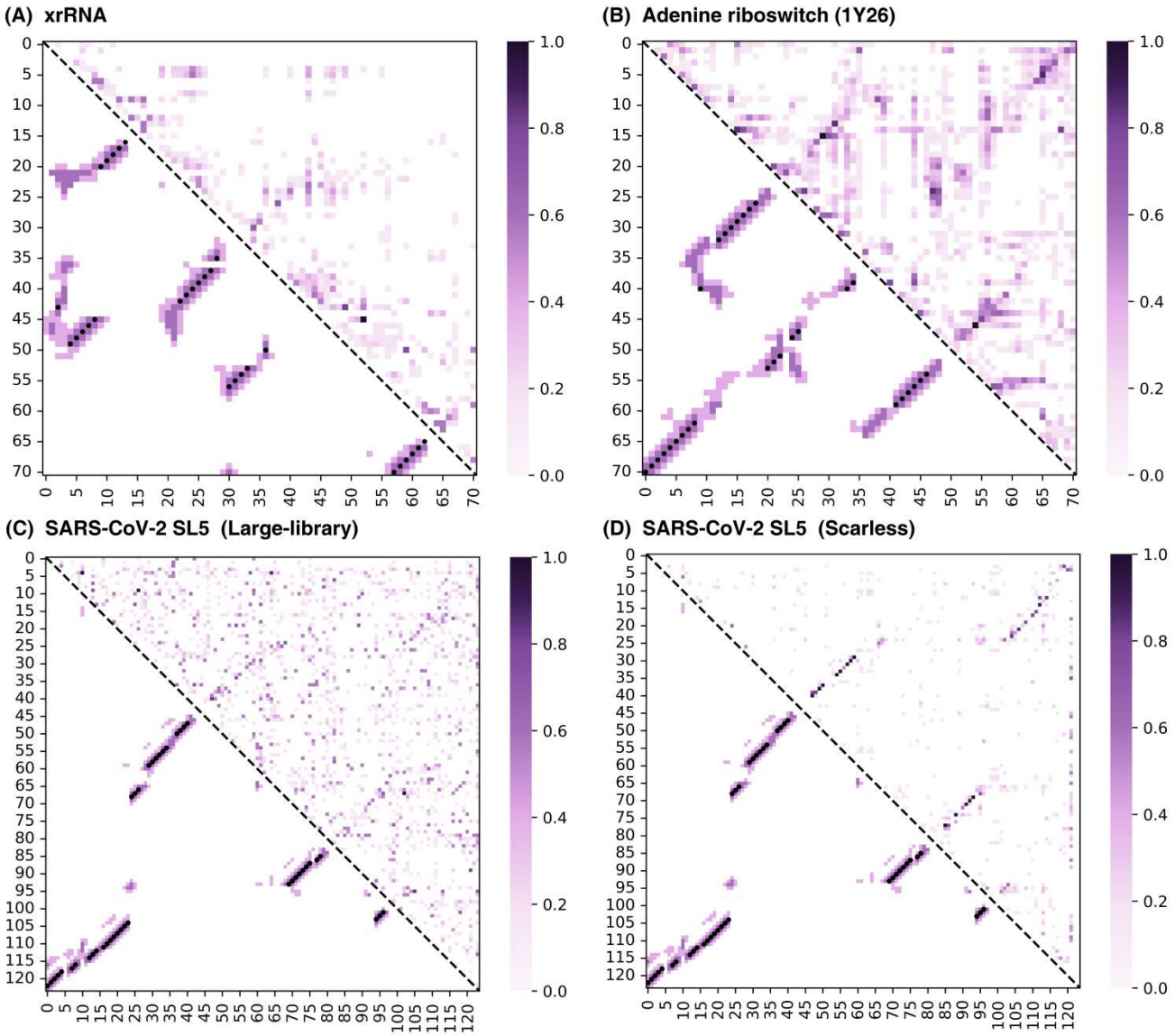



**Figure S4**. **The predicted base pairs inferred from other techniques.** The native contact maps (lower left) and the base pairs (upper right) inferred from M2-Seq (DMS probe) for xrRNA (A), Mutate-and-map (SHAPE) for adenine riboswitch (B), M2-Seq (DMS probe + Large Library) for SARS-CoV-2 SL5 (C), and M2-Seq (DMS probe + Scarless) for SARS-CoV-2 SL5 (D). In native contact maps, the native base pairs are shown in black dots, and the contacts with a 5 Å cutoff are shown in dark purple, while those with an 8 Å cutoff are shown in light purple. The corresponding *F1-scores* are shown in Supplementary Table S5, and the data used for the plots are directly sourced from the corresponding references.

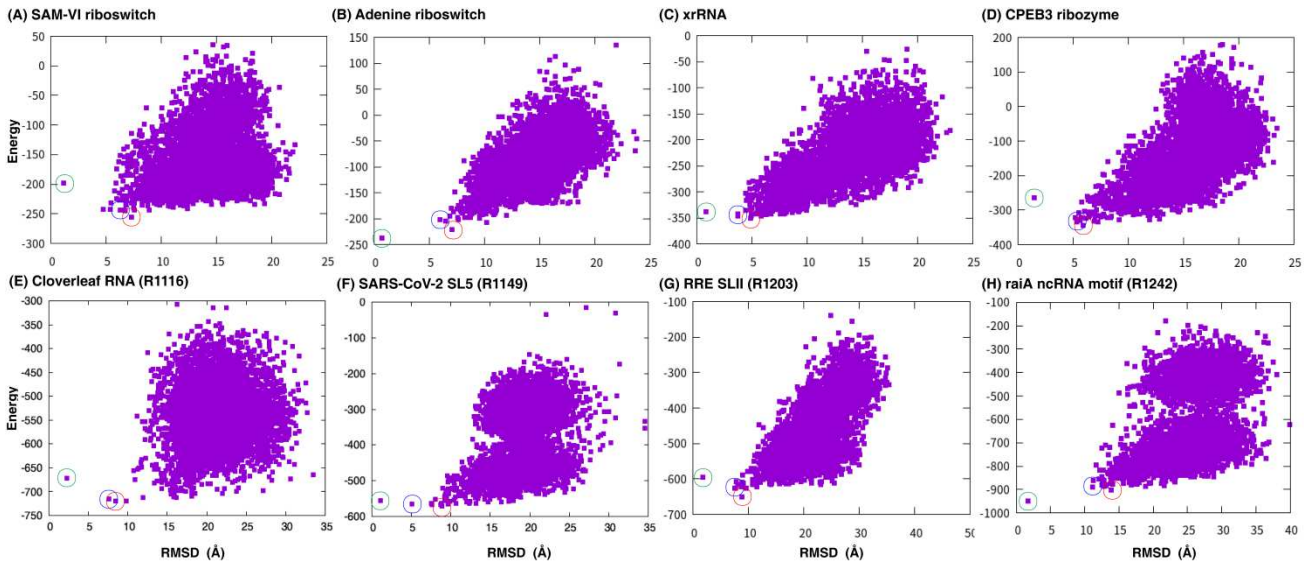



**Figure S5**. **Predicted 3D structure models evaluated by the BRiQ energy function.** BRiQ Energy versus RMSD values of the conformations sampled by BRiQ using the top 5 base-pair models inferred by CODA2+MC as the restraints for SAM-VI riboswitch (A), adenine riboswitch (B), xrRNA (C), CPEB3 ribozyme (D), Cloverleaf RNA (E), SARS-CoV-2 SL 5 (F), RRE SLII (G), and raiA (H). The points for top 1, best in top 5, and refined native structures are marked by red, blue, and green circles, respectively.

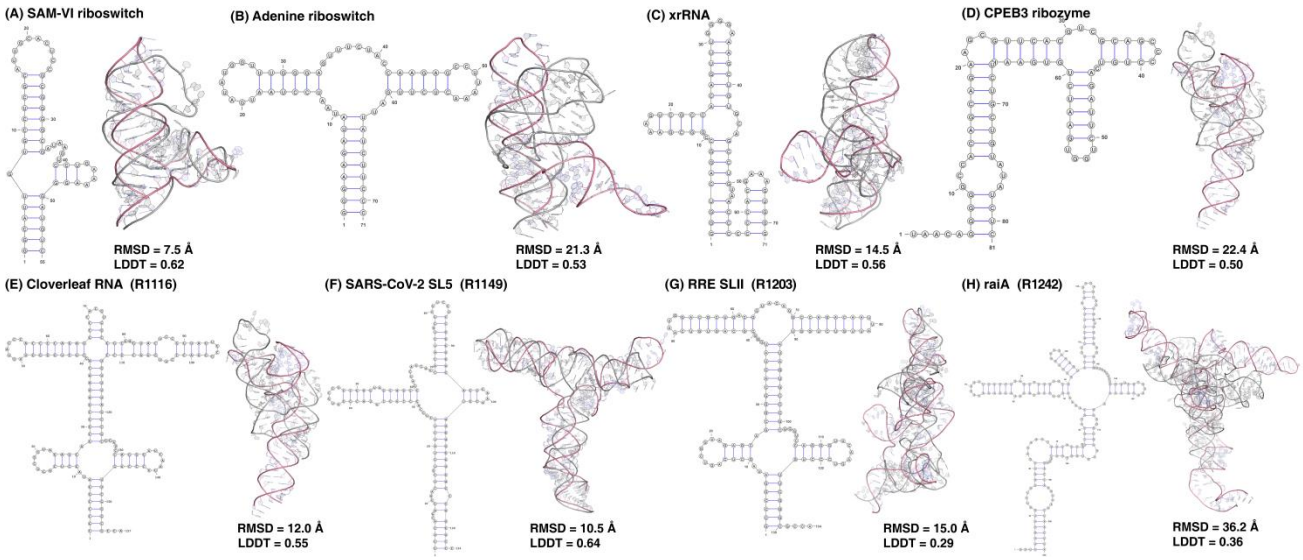



**Figure S6**. **The results from RNAfold as the baselines.** Secondary structures predicted by RNAfold as well as the top 1 of 3D structures modeled by BRiQ using the RNAfold-predicted secondary structures as restraints (grey: experimental structures; color: predicted structures.) for SAM-VI riboswitch (A), adenine riboswitch (B), xrRNA (C), CPEB3 ribozyme (D), Cloverleaf RNA (E), SARS-CoV-2 SL 5 (F), RRE SLII (G), and raiA (H).

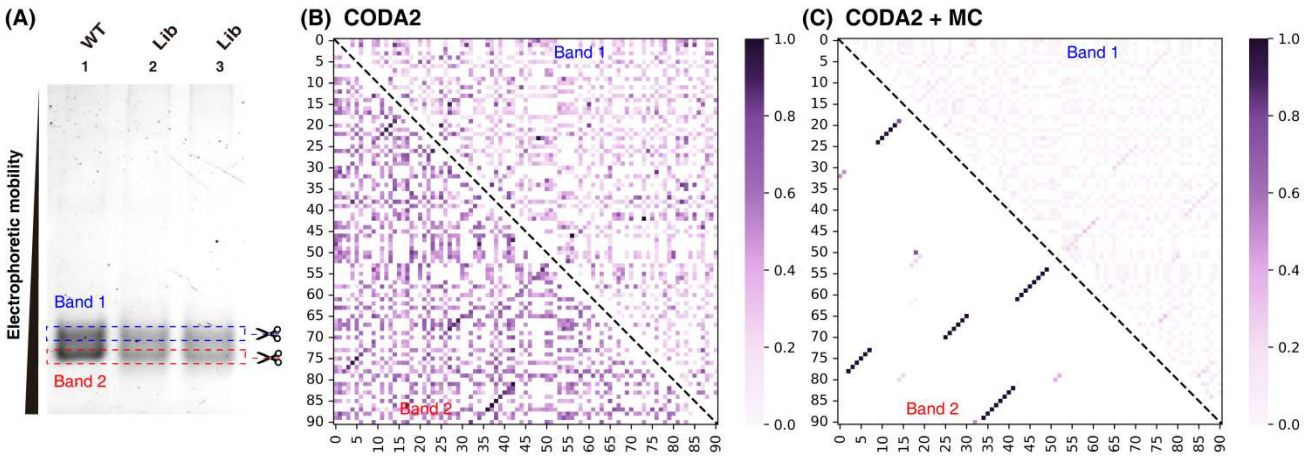



**Figure S7. RNA-MobiSeq distinguishes alternative folding states of PZ62 with high structural sensitivity.** (A) Native gel electrophoresis of PZ62 reveals two closely spaced bands (blue: Band 1; red: Band 2), indicating the presence of distinct structural conformations. Each band is separately excised and used as the reference for mutant library selection. (B) CODA2-derived contact maps for Band 1 (upper triangle) and Band 2 (lower triangle). Band 1 displays largely disordered and stemless patterns, whereas Band 2 exhibits well-defined secondary structure elements, including five stems and a prominent pseudoknot (identical to Fig. S1C). (C) Contact maps integrating CODA2 and MC simulation results, with upper and lower triangles representing Band 1 and Band 2, respectively. Although Band 1 also displays weak stem-like signals after MC integration, these are surrounded by numerous random base-pairing contacts, indicating that the stem features likely arise from thermodynamic priors used in MC modeling rather than covariation captured by CODA2. Together, these results demonstrate that RNA-MobiSeq sensitively resolves alternative RNA folding states even when their electrophoretic mobilities differ only marginally.

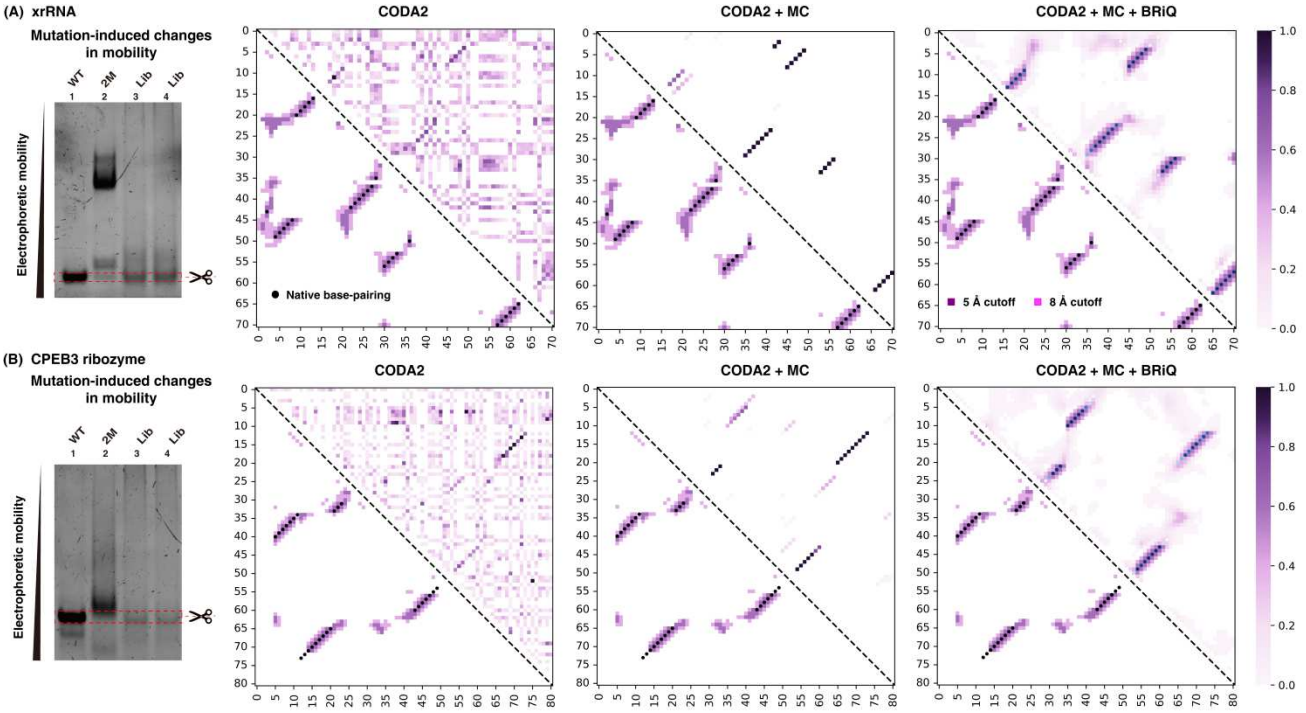



**Figure S8. Mutation induced changes in mobility, and base-pairing inferred by RNA-MobiSeq from replicate experiments of two representative RNAs.** (A, B) Mutation induced changes in mobility, and the base pair and/or distance-based contact maps (upper right) inferred from CODA2, CODA2 + MC, and CODA2 + MC + BRiQ as labelled on the top for xrRNA (A) and CPEB3 ribozyme (B).

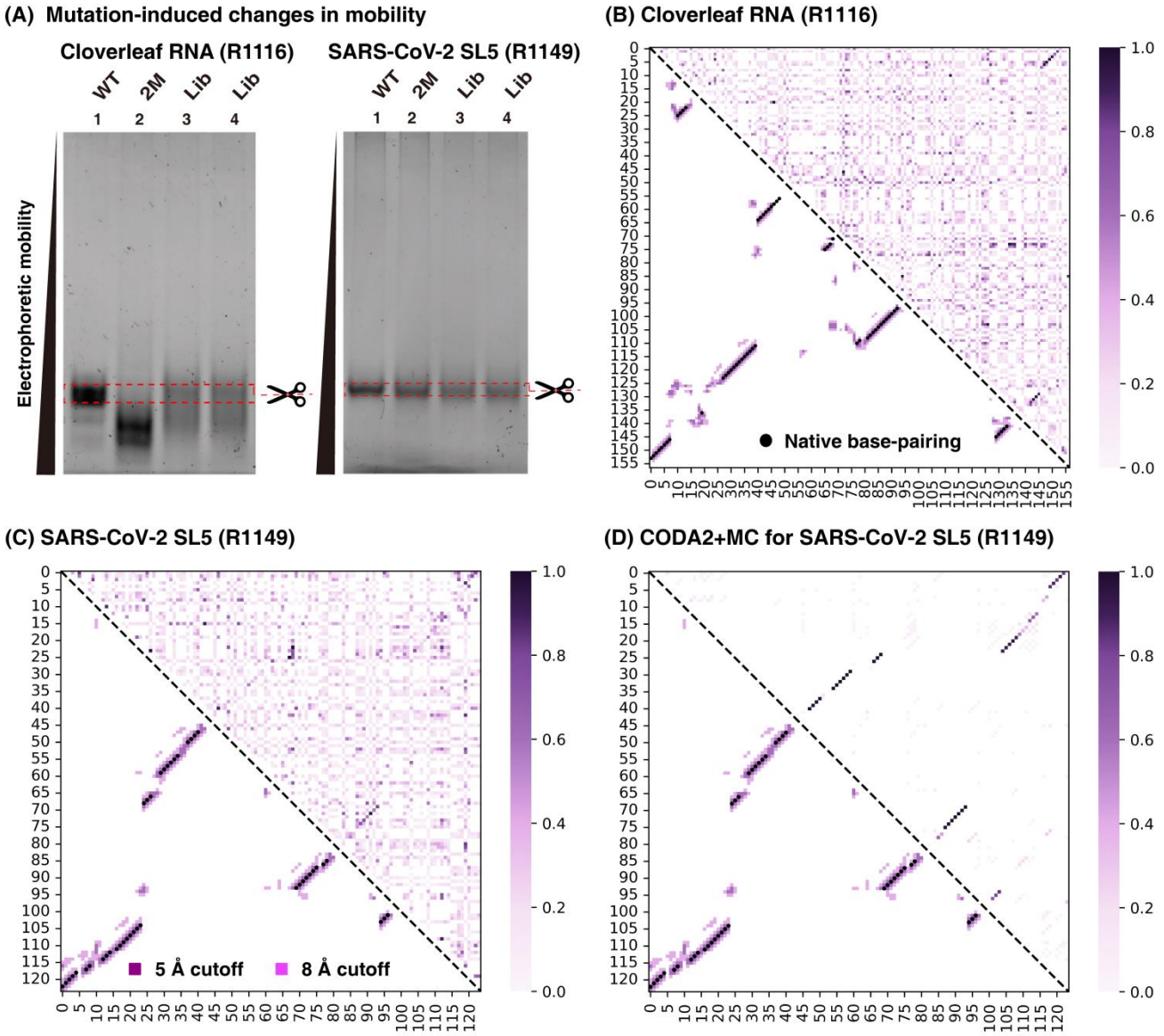



**Figure S9**. **Predicted base pairs inferred from CODA2 for two CASP 15 RNA targets with mutants synthesized using oligo pools.** (A) The mobility screening of Cloverleaf RNA (R1116; left) and SARS-CoV-2 SL5 (R1149; right. The mobility was compared between wild type (WT), double mutants (2M), and two library replicates for each RNA. (B-D) The native contact maps (lower left) and the base pairs (upper right) inferred from CODA2 for Cloverleaf RNA (B) and SARS-CoV-2 SL5 (C), as well as from CODA2+MC for SARS-CoV-2 SL5 (D). *F1-scores* inferred from CODA2 are 0.28 and 0.32 for R1116 (B) and R1149 (C), respectively. In the native contact maps, the native base pairs are shown in black dots, and the contacts with a 5 Å cutoff are shown in dark purple squares, while those with an 8 Å cutoff are shown in light purplesquares.

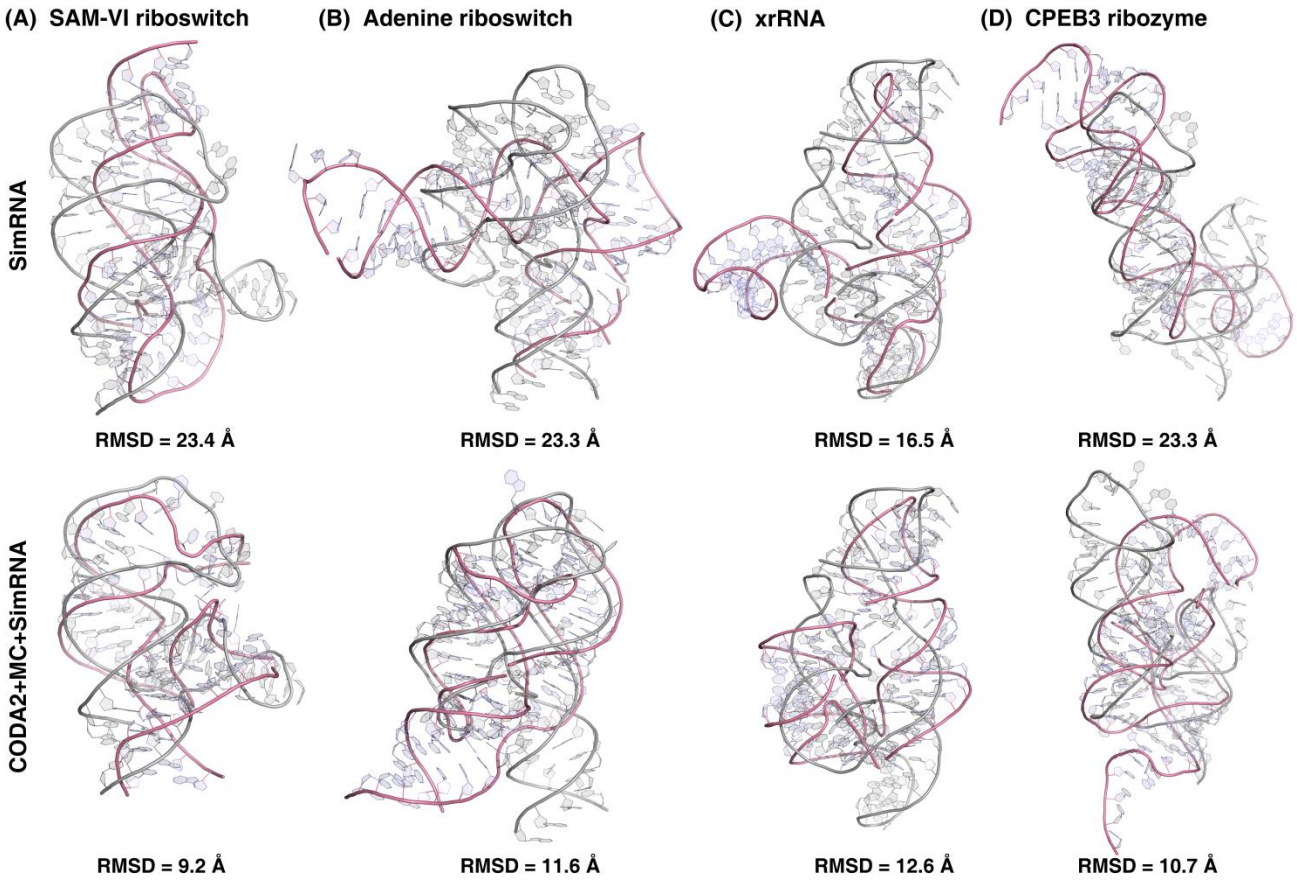



**Figure S10. The secondary structures inferred by RNA-MobiSeq can significantly enhance the accuracy of tertiary structure prediction by SimRNA**. Top: tertiary structures predicted using SimRNAweb2.0 (https://genesilico.pl/SimRNAweb/submit; with default parameters) from sequence only for four small RNAs: SAM-VI riboswitch, Adenine riboswitch, xrRNA, and CPEB3 ribozyme. Bottom: tertiary structures predicted using SimRNAweb2.0 with the top 1 secondary structure from CODA2+MC as restraints for each RNA. Grey: experimental structures. Colored: predicted structures.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MobiSeqSupplementaryMaterial.pdf
- nrreportingsummary.pdf
- nrsoftwarepolicy.pdf