

# Modeling RNA Secondary Structure with Sequence Comparison and Experimental Mapping Data

Zhen Tan,<sup>1,2</sup> Gaurav Sharma,<sup>2,3,4,\*</sup> and David H. Mathews<sup>1,2,4,\*</sup>

<sup>1</sup>Department of Biochemistry and Biophysics, <sup>2</sup>Center for RNA Biology, <sup>3</sup>Department of Electrical and Computer Engineering, and <sup>4</sup>Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York

**ABSTRACT** Secondary structure prediction is an important problem in RNA bioinformatics because knowledge of structure is critical to understanding the functions of RNA sequences. Significant improvements in prediction accuracy have recently been demonstrated through the incorporation of experimentally obtained structural information, for instance using selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) mapping. However, such mapping data is currently available only for a limited number of RNA sequences. In this article, we present a method for extending the benefit of experimental mapping data in secondary structure prediction to homologous sequences. Specifically, we propose a method for integrating experimental mapping data into a comparative sequence analysis algorithm for secondary structure prediction of multiple homologs, whereby the mapping data benefits not only the prediction for the specific sequence that was mapped but also other homologs. The proposed method is realized by modifying the TurboFold II algorithm for prediction of RNA secondary structures to utilize basepairing probabilities guided by SHAPE experimental data when such data are available. The SHAPE-mapping-guided basepairing probabilities are obtained using the RSample method. Results demonstrate that the SHAPE mapping data for a sequence improves structure prediction accuracy of other homologous sequences beyond the accuracy obtained by sequence comparison alone (TurboFold II). The updated version of TurboFold II is freely available as part of the RNAstructure software package.

## INTRODUCTION

RNA functions in diverse cellular activities; it is a carrier of genetic information in transcription (1), a regulator of gene expression (2), and a catalyst (3). These cellular functions depend on the structure of RNA (4). Therefore, accurate predictions for the secondary structure, i.e., canonical basepairings between nucleotides, are critical for understanding and proposing hypotheses related to RNA functions. A commonly used approach is to predict secondary structures based on folding thermodynamics (5,6).

To achieve greater prediction accuracy, several thermodynamics-based methods incorporate experimental data derived from chemical probing to guide RNA secondary structure prediction (7–17). One mapping method, selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), provides quantitative reactivity at each nucleotide to the SHAPE reagent, which measures the nucleotide flexibility (18,19). Because basepaired nucleotides are structurally restricted, high SHAPE reactivity is generally

associated with not being canonically basepaired (20). SHAPE data can be collected with high-throughput sequencing (21–23) and can also be obtained in vivo (24–26).

RSample (Spasic, S.M. Assmann, P.C. Bevilacqua, D.H.M., unpublished data) models RNA secondary structure using SHAPE data. It focuses on matching structure models to the mapping data rather than directly integrating data into the model. In this way, it can model folding ensembles of multiple structures. A nucleotide-level comparison between experimental mapping data and modeled mapping data is used to guide a single refinement of a stochastic sample. The sample is then clustered to predict sets of structure models. The single structure prediction accuracy of RSample is similar to leading methods (>80% of predicted pairs in the accepted structure) (12), and RSample is able to estimate the population of multiple structures in the folding ensemble (27).

Another approach to improving secondary structure prediction accuracy is to use multiple homologous sequences to identify conserved basepairs (5,28–30). One method, TurboFold II (31; Z.T., Y. Fu, G. Sharma, D.H.M., unpublished data), iteratively refines basepairing probabilities for each sequence in a set of homologs by comparing the predicted basepairing probabilities across the set of

Submitted March 1, 2017, and accepted for publication June 19, 2017.

\*Correspondence: [david\\_mathews@urmc.rochester.edu](mailto:david_mathews@urmc.rochester.edu) or [gaurav.sharma@rochester.edu](mailto:gaurav.sharma@rochester.edu)

Editor: Tamar Schlick.

<http://dx.doi.org/10.1016/j.bpj.2017.06.039>

© 2017 Biophysical Society.

homologs. Additionally, nucleotide alignment probabilities in pairwise alignments, estimated using a hidden Markov model (HMM) (32), are iteratively improved using information from estimated secondary structures (33). After the iterative updates, structures are predicted using the maximum expected accuracy algorithm (34–36) and a multiple sequence alignment is estimated using a probabilistic consistency transformation (36) and progressive alignment.

An open problem in the field is the integration of both structure mapping data and comparative data to improve secondary structure prediction accuracy. Prior work focused on the case where SHAPE data is available for all homologous sequences (37). For this situation, a multiple sequence alignment was first created by also including SHAPE data in pairwise global alignment. Then the RNAalifold method (38) was used to predict a consensus structure that is conserved given the fixed input alignment, using pseudo free energies to incorporate the SHAPE information (7). This article addresses the problem of predicting conserved secondary structures when SHAPE mapping is only available for one homolog. This use case is expected to be increasingly common as SHAPE is performed in vivo across transcriptomes. The method reported in this article is the integration of RSample into TurboFold II. In the resulting method, SHAPE-guided structure prediction and prediction of conserved structures act synergistically to improve secondary structure prediction accuracy, even for sequences for which SHAPE mapping was not performed. Results demonstrate that the SHAPE mapping data for a sequence improves structure prediction accuracy of other homologous sequences beyond the accuracy obtained by sequence comparison alone (TurboFold II).

## METHODS

Fig. 1 illustrates the proposed new version of TurboFold II that uses available SHAPE mapping data for one or more of the RNA sequence homologs to improve structure prediction for the sequences without SHAPE data. The input to TurboFold II is a set of homologous sequences and the outputs are the predicted secondary structures for each sequence and a multiple sequence alignment (31). To incorporate experimental mapping data into the predictions, the proposed approach makes use of RSample. Specifically, as shown in Fig. 1, within the TurboFold II iterations, RSample is used to refine estimated basepairing probabilities for sequences with SHAPE data and these estimated basepairing probabilities are incorporated in the iterations. As shown via the dashed lines in Fig. 1, in subsequent TurboFold II iterations, the incorporated SHAPE information propagates to other homologous sequences and thereby improves the prediction of structure for these sequences, in addition to improving structure prediction for the sequence with which the SHAPE data is affiliated. The major individual steps in the proposed approach are outlined next.

### SHAPE-guided computation of basepairing probabilities using RSample

RSample first generates a stochastic sample (39) using a secondary structure partition function calculation (40). Then SHAPE reactivities are esti-

mated for each nucleotide in each structure based on the status of the nucleotide: unpaired, paired at the last position of a helix, or paired in the interior of a helix. SHAPE reactivities are drawn from distributions composed of a database of 16 known secondary structures with experimentally measured SHAPE reactivities (12). The estimated SHAPE reactivity for a nucleotide is then the mean reactivity across all structures. The stochastic sampling is then repeated, where the partition function is reestimated so that the estimated SHAPE reactivities better match the experimental SHAPE mapping data. The free energy change term introduced to the partition function is

$$\Delta G_{\text{bonus},i} = 0.5 \times \ln \left( \frac{R_{\text{exp}_i} + 1.1}{R_{\text{calc}_i} + 1.1} \right), \quad (1)$$

where  $R_{\text{exp}_i}$  and  $R_{\text{calc}_i}$  are experimentally measured reactivities and estimated reactivities of nucleotide  $i$ . This functional form was chosen so that the free energy of basepair stacking is only altered for nucleotides for which the originally estimated SHAPE reactivity does not match the experiment. The constants 0.5 and 1.1 in the equation were obtained (data not shown) via a grid search as the parameters that maximized structure prediction accuracy. The free energy bonus  $\Delta G_{\text{bonus},i}$  is then applied for each basepair stack involving nucleotide  $i$ . This approach focuses on matching the experimentally measured SHAPE reactivity.

## Incorporation of RSample into TurboFold II

TurboFold II is a method to predict secondary structures for multiple RNA homologs and multiple sequence alignments. TurboFold II iteratively estimates basepairing probabilities for each sequence using intrinsic information and extrinsic information for sequence folding. Intrinsic information is derived from the thermodynamic model, which used the latest set of nearest-neighbor thermodynamic parameters (11,41). Extrinsic information is a proclivity for basepairing inferred from the basepairing probabilities of other homologous sequences, mapped to the sequence of interest by the posterior probabilities of nucleotide coincidence of the other homologs to the sequence (32). The posterior coincidence probabilities can be obtained with a HMM for pairwise alignments (42). The estimated basepairing probabilities can be used to predict secondary structure using the maximum expected accuracy (MEA) algorithm (34,35,43) or the ProbKnot method (44).

RSample is integrated into TurboFold II to estimate basepairing probabilities for homologous sequences with available SHAPE mapping data on one of the homologs. The integrated algorithm uses nine steps illustrated in Fig. 1.

We adapt the description focusing particularly on the new elements introduced in this article.

Step 1 computes pairwise posterior coincidence probabilities using an HMM. Pairwise posterior coincidence probabilities are estimated for all pairs of sequences with an HMM as implemented by Harman et al. (32). Using the forward-backward algorithm, matrices of posterior coincidence probabilities for two nucleotides (one from each sequence) are computed. Details can be found in Harman et al. (32).

Step 2 computes basepairing probabilities of all sequences using the partition function method in RNAstructure (40).

Steps 3–5 are only performed for sequences for which there is SHAPE mapping data.

Step 3 generates an ensemble of  $N_s = 10,000$  structures by stochastic sampling for sequences with input SHAPE reactivity.

Step 4 estimates the SHAPE reactivity for each nucleotide based on the sample. The SHAPE reactivities are assigned to each nucleotide at each structure in the sample according to the distributions for three different local structures: unpaired, paired at a helix end, or paired in the interior of a helix. The SHAPE reactivity for each nucleotide is the arithmetic mean across structures in the sample. Because the size of ensemble is large, the variance between samples is relatively low.

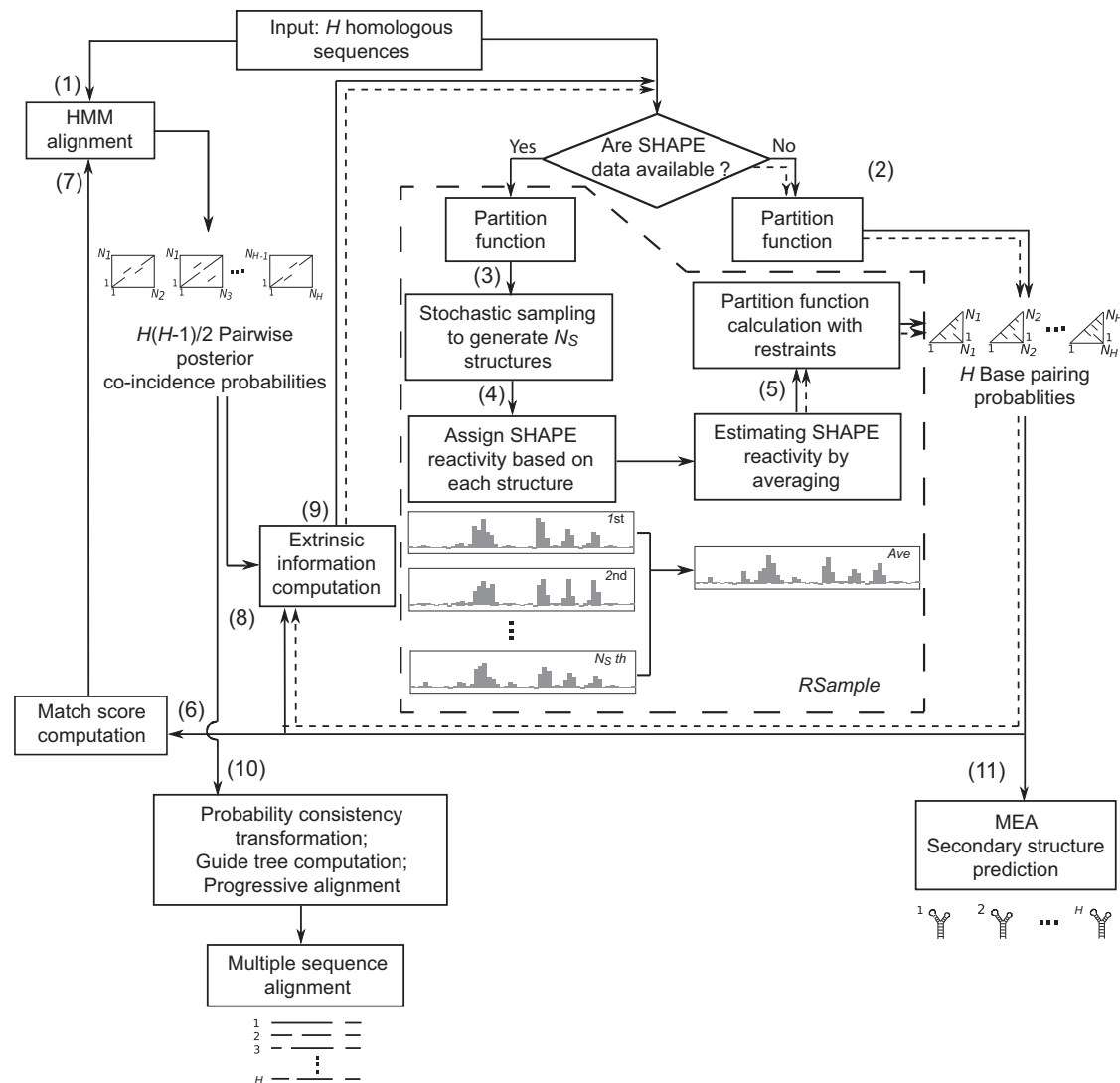


FIGURE 1 Flowchart for TurboFold II with incorporation of SHAPE mapping data for one or more sequences. The input is a set of  $H$  homologous RNA sequences and the outputs are the predicted secondary structures for each sequence and the predicted multiple sequence alignment. Steps 1–11 are described in Materials and Methods. The dashed arrow lines show the flow of SHAPE information and illustrate how, through the iterations, the SHAPE information contributes not only to the structure prediction for sequences with SHAPE data but also to the structure prediction for other sequences. Steps 3–5 in the dashed box show the processing for the sequences with SHAPE mapping data using RSample.

Step 5 recalculates the partition function using the free energy change term (in Eq. 1) to predict basepairing probability for the sequence with input SHAPE reactivities. Nucleotides with higher or lower estimated SHAPE reactivity than that measured by experiment are restrained with a lower or higher propensity to basepair, respectively. Nucleotides with consistent estimated and experimental SHAPE reactivity receive no restraint.

Step 6 calculates match scores that encourage alignment between nucleotide positions where both nucleotides are upstream paired, downstream paired, or unpaired. The match score was first proposed in PMcomp (33), and is utilized in TurboFold II as a prior for recalculating posterior coincidence probability in next step via the HMM pair alignment algorithm. For the  $m$ th sequence, based on estimated basepairing probabilities between all pairs of nucleotide positions obtained from the partition function calculation, for a nucleotide at position  $i$ , the estimated probability of downstream pairing is  $P_{<}^m(i) = \sum_{j>i} P_{ij}^m$ , of upstream pairing is  $P_{>}^m(i) = \sum_{j<i} P_{ij}^m$ , and of being unpaired is  $P_{\circ}^m(i) = 1 - P_{<}^m(i) - P_{>}^m(i)$ .

The match score between nucleotides  $i$  and  $k$  in sequences  $m$  and  $n$ , respectively, is formulated as

$$\rho(i, k) = \left( \sqrt{P_{<}^m(i)P_{<}^n(k)} + \sqrt{P_{>}^m(i)P_{>}^n(k)} \right) + 0.8 \\ \times \left( \sqrt{P_{\circ}^m(i)P_{\circ}^n(k)} \right) + 0.5. \quad (2)$$

For sequences without SHAPE mapping data, the basepairing probabilities from Step 2 are utilized for the computation of match scores, whereas for sequences with SHAPE mapping data, the basepairing probabilities from Step 5 are used in the computation of the match scores.

Step 7 reestimates the posterior coincidence probability. Information from prior iterations is utilized to reestimate alignment posterior probabilities and basepairing probabilities for secondary structures. The iterative reestimation of alignment posterior probabilities is introduced (TurboFold

II) and uses the standard HMM alignment model (42), but with the match score of Eq. 3 incorporated as a prior.

Step 8 calculates extrinsic information for each sequence by combining basepairing probabilities from other input sequences using posterior coincidence probabilities:

$$P^{(n \rightarrow m)}(i, j) = \sum \begin{cases} \sum_{\substack{k, l \\ 1 \leq k < l \leq N_n \\ k \in C_i^{m, n} \\ l \in C_j^{m, n}}} \text{Prob}_{bp}(k, l) \times P^{(m, n)}(i \sim k) \times P^{(m, n)}(j \sim l) \times (H - 1) \times \lambda & \text{(if sequence } n \text{ is with SHAPE)} \\ \sum_{\substack{k, l \\ 1 \leq k < l \leq N_n \\ k \in C_i^{m, n} \\ l \in C_j^{m, n}}} \text{Prob}_{bp}(k, l) \times P^{(m, n)}(i \sim k) \times P^{(m, n)}(j \sim l) \times (1 - \psi_{m, n}) & \text{(otherwise),} \end{cases} \quad (3)$$

where  $P^{(n \rightarrow m)}$  denotes the extrinsic information for sequence  $m$  inferred from sequence  $n$ .  $N_n$  indicates the length of sequence  $n$ . The notations  $C_i^{m, n}$  and  $C_j^{m, n}$  denote the sets of indices for which posterior coincidence alignment probabilities  $P^{(m, n)}(i \sim k)$  and  $P^{(m, n)}(j \sim l)$ , respectively, exceed a predetermined threshold below which values are considered 0 for computational simplification.  $\text{Prob}_{bp}(k, l)$  denotes the (estimated) basepairing probability between nucleotide  $k$  and nucleotide  $l$  within a sequence. The value “ $i \sim k$ ” indicates the alignment between indices  $i$  and  $k$  in two sequences.  $H$  is the number of homologous sequences. To keep the ratio of extrinsic information from sequence  $n$  to every other sequence constant, the extrinsic information term for sequence  $n$  is multiplied by  $H-1$  if sequence  $n$  has SHAPE data. This ensures that more extrinsic information is used from sequences with SHAPE data than from sequences without SHAPE data.  $\lambda$  is a parameter, optimized based on training. The factor  $(1 - \psi_{m, n})$  weights the contribution of sequence  $n$  to the extrinsic information for sequence  $m$  using the sequence identity,  $\psi_{m, n}$ , for sequences  $m$  and  $n$  computed from an HMM alignment. This term is only used when sequence  $n$  does not have associated SHAPE mapping data. Because of the factor  $(1 - \psi_{m, n})$ , sequences that are highly similar to sequence  $m$  have a lower contribution to extrinsic information than those with lower similarities. The extrinsic information is calculated from basepairing proclivity for each sequence as inferred from every other sequence pairwise. Because the sequence with SHAPE reactivities is presumed to have more accurate estimates of basepairing probabilities, the basepairing proclivities from the sequence with SHAPE reactivities to sequences without SHAPE reactivities are assigned a different, adjustable weighting ( $\lambda$ ). The basepairing proclivities for sequences without SHAPE data and from other sequences to the sequence with SHAPE data are computed in an identical fashion to the TurboFold II algorithm.

Step 9 updates the basepairing probability by recomputing the partition function for each sequence with the addition of extrinsic information. The extrinsic information is incorporated as a pseudo free energy term in the partition function calculation for each sequence. A detailed description is in Harmanci et al. (31).

Steps 2–9 form a loop that is iterated through three times, which is shown to be optimal in Harmanci et al. (31).

Steps 10 and 11 perform progressive alignment and predict final secondary structures, respectively. In Step 10, the posterior coincidence

probabilities obtained with the updated match scores via Step 6 are used to calculate a multiple sequence alignment. A probabilistic consistency transformation, as described in ProbCons (36), is used to refine alignment probabilities based on three-way alignment consistency of pairwise HMM posterior probabilities. Refined alignments are

further predicted hierarchically based on a guide tree, as described in ProbCons (36).

In Step 11, the structures are predicted by the MEA algorithm. Given the basepair probabilities  $P^m(i, j)$  for structure  $s_m$  of sequence  $m$ , the MEA structure is defined as

$$S_m^* = \underset{S_m}{\text{argmax}} \left\{ \sum_{(i, j) \in S_m} 2 \cdot P^m(i, j) + \sum_{\substack{\forall i; \\ i \text{ unpaired in } S_m}} P^m(i) \right\}, \quad (4)$$

where  $P^m(i)$  is the probability that nucleotide position  $i$  is not basepaired, computed as

$$P^m(i) = 1 - \sum_{j=i+1}^{N_m} P^m(i, j) - \sum_{j=1}^{i-1} P^m(j, i), \quad (5)$$

and where  $N_m$  is the length of sequence  $m$ . The MEA structure is obtained with a dynamic programming algorithm, as described in Harmanci et al. (31).

## Parameter optimization

To train the parameter  $\lambda$  corresponding to the weighting of the extrinsic information term in Eq. 3, 20 groups of input sequences formed by 10 homologous sequences (including the sequence with SHAPE data) were randomly chosen from the small subunit ribosomal RNA in the RNAStralign database. The range for parameter  $\lambda$  was from 0 to 2.0 (with samples at 0, 0.02, 0.1, 0.2, 0.4, 1.0, 1.6, and 2.0). The resulting optimal parameter ( $\lambda = 1.0$ ) was then used as the default for the method. The geometric mean of sensitivity and PPV was used as the accuracy metric for optimizing the parameter  $\lambda$ , and the values of this metric over the training set are given in the Supporting Material (Fig. S15).



## Benchmarks

For benchmarking, groups of sequence homologs were selected from several families based on the selection criterion that SHAPE data were available for a sequence in the family (12). Hepatitis C virus (HCV) IRES domain, TPP riboswitch, cyclic-di-GMP riboswitch, SAM I riboswitch, M-box riboswitch, and Lysine riboswitch RNA sequences were randomly selected from the Rfam database (45). tRNA, 5S ribosomal RNA, and group I intron sequences were selected from the RNAStralign database (<http://rna.urmc.rochester.edu/RNAStralign.tar.gz>). 23S rRNA sequences were selected from the Comparative RNA web site and project (<http://www.rna.icmb.utexas.edu/>). Specifically, 20 groups of 4-, 9-, or 19-sequence homologs were selected from each of the RNA family. All methods were benchmarked on the same groups of sequences. Detailed information of selected sequences is in Tables S1 and S2. For comparison, a single sequence prediction accuracy was also computed as the average of the accuracies for each homolog in the set of sequences for predictions obtained using the MaxExpect (maximum expected accuracy) method from RNAstructure 5.7.

## Scoring of prediction accuracy

The F1 score, which is the harmonic mean of sensitivity and PPV, is used in the structure-prediction benchmark. The F1 score is computed as

$$F1 = \frac{2 \times \text{Sensitivity} \times \text{PPV}}{\text{Sensitivity} + \text{PPV}}. \quad (6)$$

Sensitivity is the fraction of basepairs from the Rfam database that are correctly predicted. PPV is the fraction of predicted basepairs that are correct, i.e., included in the Rfam database.

Predicted basepairs are considered correct if a nucleotide on either the 5'- or 3'-position of the helix is off by one position compared to the standard (13,46). For instance, a predicted basepair ( $i, j$ ) is correct if basepair ( $i, j$ ), or ( $i \pm 1, j$ ), or ( $i, j \pm 1$ ) exists in the database. This is important because of uncertainty in the determination of secondary structure by comparative analysis (47) and also because of thermodynamic fluctuations of local structures (48,49).

## Significance testing

To assess the statistical significance of the differences in F1 score, sensitivity, and PPV, paired *t*-tests were performed using R 3.0.2 (50) between TurboFold II with SHAPE data and each of the other methods (51). Alpha, the type I error rate, was set to 0.05. The figures summarizing the benchmarking results are annotated to indicate the results of the significance tests.

## Alternative methods

Although no previous work has been reported on using SHAPE data for one homolog in the prediction of structures for other homologs, the RNAalifold (38,52) method can be used for this purpose and it is therefore used for comparison. For RNAalifold, the SHAPE reactivity data is converted to per-nucleotide pseudo free energies that are then applied for each basepair stack including a nucleotide. A log-linear fit based on Deigan et al. (7) is used to convert reactivities into pseudo free energies. The RNAalifold method does not compute an alignment and requires an input multiple sequence alignment. Input alignments for RNAalifold (2.2.5) were generated using ClustalW (2.1) (38,53). Default options and parameters were used for these programs in the benchmarking.

## RESULTS

The new version of TurboFold II, capable of incorporating SHAPE data, was benchmarked for structure prediction accuracy using RNA families, where one sequence in each family has measured SHAPE reactivity (12). The method was compared with RNAalifold (38), RSample, and MaxExpect (35). RNAalifold is a method for predicting consensus structures for multiple homologs. It was previously adapted for using SHAPE data, and was benchmarked for cases when all sequences had SHAPE mapping data (37). RSample is run for the single sequences with SHAPE data available. MaxExpect is the single sequence maximum expected accuracy method, and maximum expected accuracy is used to generate the predicted structures from predicted basepairing probabilities with TurboFold. The accuracy results are represented in Figs. 2 and S1–S11; Tables S4 and S5.

Fig. 2 shows the average structure prediction accuracy for the sequences without SHAPE data. The results demonstrate that the majority of RNA families (tRNA, 5S rRNA, hepatitis C virus IRES, group I intron, lysine riboswitch, SAM I riboswitch, cyclic-di-GMP riboswitch, and 23S rRNA) have significantly ( $p < 0.05$ ) better structure prediction accuracy when SHAPE is used in the calculation than when SHAPE data is not used. This shows that SHAPE data for a single sequence can inform the structure modeling for homologous sequences. However, for the M-box riboswitch and TPP riboswitch, the accuracies are not significantly improved by having SHAPE data. For the sequences without SHAPE data, the new version of TurboFold II performed better than RNAalifold using SHAPE data and MaxExpect. Fig. S12 shows that much of the improvement in accuracy is for sequences that were relatively poorly predicted in the absence of SHAPE data. The accuracy performance for those sequences is rescued by having SHAPE information for a homologous sequence.

It is observed that structure prediction accuracies by TurboFold II using SHAPE data across sizes of sequence groups are scarcely changed (from 5 to 20 sequences). The relationship between structure prediction accuracies and sequence lengths is also weak (Tables S1 and S2). For the 23S rRNA family, which has the longest average sequence length (~2900 nucleotides), all methods, except single-sequence MaxExpect, perform well. On the RNA families with sequence lengths shorter than 200 nucleotides, TurboFold II + SHAPE improves structure predictions for tRNA, 5S, lysine riboswitch, and cyclic-di-GMP riboswitch, but does not improve structure predictions for M-box riboswitch and TPP riboswitch.

For the one sequence with SHAPE mapping data in each RNA family, the results show that the majority of RNA families (5S rRNA, HCV IRES domain, group I intron, TPP riboswitch, and 23S rRNA) have significantly ( $p < 0.05$ ) improved prediction accuracy when SHAPE data are used

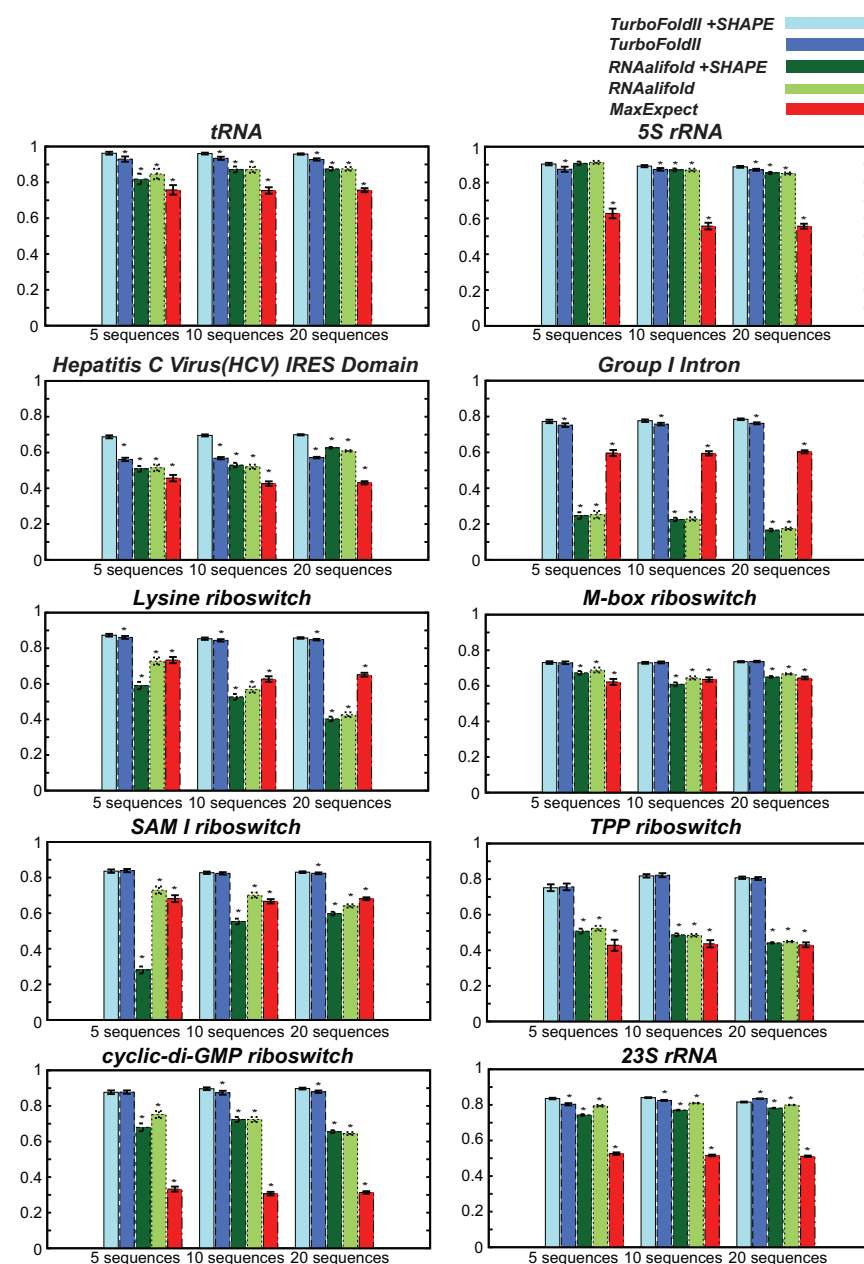


FIGURE 2 Average F1 score of structure predictions of the sequences that did not have SHAPE mapping data. Given here is the average F1 score of structure predictions obtained by running the methods with 5-, 10-, or 20-input sequences on tRNA, 5S rRNA, hepatitis C virus IRES domain, group I intron, lysine riboswitch, M-box riboswitch, SAM I riboswitch, TPP riboswitch, cyclic-di-GMP riboswitch, and 23S rRNA test datasets. Standard errors of the mean are shown by error bars. The star (\*) above the bar for a method indicates that the difference in F1 score between the method and the new TurboFold II is statistically significant, as determined by paired *t*-tests (51).

than when SHAPE data are not used (Fig. S1 and Table S4). For tRNA, the lysine riboswitch, and the M-box riboswitch families, the accuracy performances are the same. In the SAM I riboswitch and the cyclic-di-GMP riboswitch families, the accuracies decreased when SHAPE data are used. In tRNA, 5S rRNA, group I intron, lysine riboswitch, SAM I riboswitch, TPP riboswitch, and 23S rRNA families, the new version of TurboFold II performed better than RSample. Only in the hepatitis C virus IRES domain and cyclic-di-GMP riboswitch families, the new version of TurboFold II performed worse than RSample. The TurboFold II+SHAPE performed better than RNAalifold using SHAPE data on every family and performed better than MaxExpect on a majority of families (except the cy-

clic-di-GMP riboswitch and the M-box riboswitch) using SHAPE data.

The alignment predictions by TurboFold II with and without SHAPE (Fig. S13) are compared with the predicted alignment by ClustalW (53), a method that is based on pairwise dynamic programming alignments, which is the input alignment for RNAalifold. Because the Rfam database alignments do not include the sequence with SHAPE data for all of the families, the alignment accuracy is assessed only over the sequences without SHAPE data within each family of homologs. With the exception of the 5S rRNA and the hepatitis C virus IRES domain, TurboFold II with SHAPE had higher sensitivity and PPV compared to ClustalW. Using SHAPE data on one sequence in each

RNA family also significantly improved the alignment accuracy of other homologs without SHAPE in a majority of RNA families (group I intron, lysine riboswitch, M-box riboswitch, SAM I riboswitch, TPP riboswitch, and cyclic-di-GMP riboswitch).

## DISCUSSION

Secondary structure models are important for understanding the functions of the RNA structure (54). Using SHAPE data was shown to improve structure prediction accuracy significantly for single sequence secondary structure predictions (7,12). In this work, it is demonstrated that the SHAPE data can inform the folding of other homologs by combining information from sequence comparison of RNA homologs. In particular, it is shown that given SHAPE data for one sequence out of the multiple sequences used in secondary structure prediction by comparative analysis, TurboFold II + SHAPE can substantially improve the structure prediction accuracies of the sequences that did not have SHAPE mapping data.

One of the reasons for the improvements of the structure prediction accuracies of homologs without SHAPE is the more accurate prediction of the structure of the sequence with SHAPE reactivity. In three RNA families (5S rRNA, HCV IRES, and group I intron), TurboFold II improved the average structure accuracy of both the sequences with and without SHAPE (Fig. S1). The more accurate structural information from the sequence with SHAPE is transmitted to its homologs through the extrinsic information calculation. Due to the specially designed extrinsic information calculation from the sequence with SHAPE to other ( $H-1$  total) homologs by introducing the factor ( $H-1$ ), which ensures that the fraction of extrinsic information provided by sequences with SHAPE is high compared to other homologs, the structure prediction of homologs is improved.

To take the advantage of SHAPE data on one of the homologs, the new method ignores pairwise sequence identity during the calculation of extrinsic information from the

sequence with SHAPE to other sequences. To understand the nature of improvements in structure prediction accuracies of sequences without SHAPE, the relationship between structure prediction accuracy and sequence identity is studied (Fig. S14). Sequence identity is defined as the ratio of the number of columns with same pairwise aligned nucleotides at the output alignment between the sequence with SHAPE and other homologs from the TurboFold II + SHAPE method. One observed trend is that the sequences with more accurately predicted structure (higher F1 score) generally with had higher sequence identity to the sequence with SHAPE. Moreover, the F1 score improvements were distributed in a roughly Gaussian shape along the sequence identity (Fig. S14). For the sequences with relatively high sequence identity, the room to improve accuracy was limited. The Gaussian shape is also partially due to the effects of improvements in structure prediction because of a more accurate alignment. This is observed in some of the RNA families (tRNA, group I intron, lysine riboswitch, and SAM I riboswitch) (Fig. S13). The 5S rRNA, hepatitis C virus IRES domain, and cyclic-di-GMP riboswitch RNA families showed improvements on structure prediction accuracy although little or no improvement on alignment prediction accuracy, because the alignment accuracies of these RNA families were already relatively high ( $\sim 90\%$  in sensitivity and PPV).

The other reason for the improvements of the structure prediction accuracies of homologs without SHAPE is the more accurate coincidence probability as compared to the case without SHAPE data on any of the input sequences. The coincidence is important to map the basepairing probabilities of other homologous sequences to the sequence of interest and it is also helpful to estimate the final multiple sequence alignment (Fig. S13).

One remaining challenge of structure prediction using experimental probing data on one of the homologs is the difficulty to determine the balance of information from thermodynamics of the sequence and extrinsic information from the sequence using experimental data. In Fig. 3, an example from the TPP riboswitch family shows that the structure of

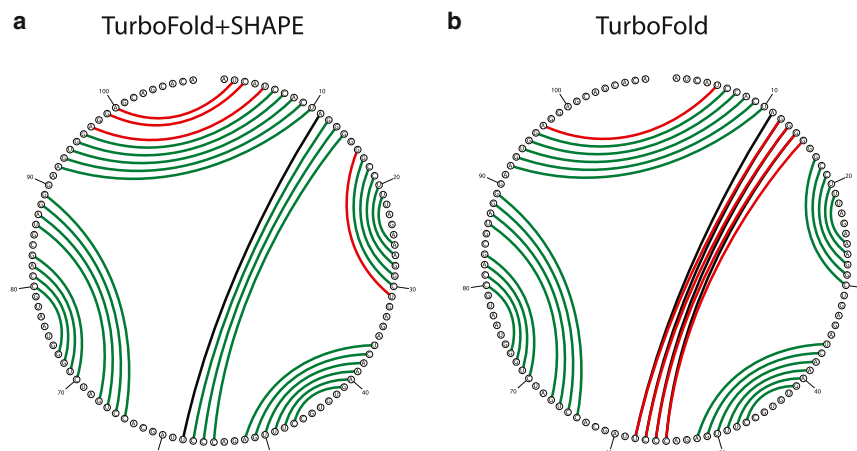


FIGURE 3 Representative secondary structure prediction for TPP riboswitch (BA000043) with (a) and without (b) SHAPE data on a homologous RNA. Basepair predictions are illustrated by colored lines (green, red, and black denoting correct, incorrect, and missing basepairs, respectively) on circle plots. The circular plots were generated using the CircleCompare program in RNAstructure (55).

one homologous sequence BA000043 was incorrectly predicted to form three extra basepairs between 5' and 3' ends when SHAPE was used as compared to when SHAPE was not used, although the longer helix contributes to a more stable structure.

RNAalifold showed lower accuracies for predicted structures than those of TurboFold II + SHAPE in most of the RNA families. A contributing factor to this inaccuracy was the lower accuracy of the input sequence alignment (Fig. S13). Although pseudo free energies obtained from the SHAPE reactivity data at nucleotides might be helpful for estimating the structure, an inaccurate alignment between the sequence with SHAPE data and homologs can disturb the consensus structure for the set of aligned sequences and can cause loss of basepairs in the consensus structure. For the group I intron, lysine riboswitch, SAM I riboswitch, TPP riboswitch, and cyclic-di-GMP riboswitch RNA families, the sensitivity and PPV of the predicted ClustalW alignment for sequences without SHAPE are ~10% lower than those of TurboFold II + SHAPE and the F1 score of structure prediction on these RNA families is ~20% lower than TurboFold II + SHAPE.

Another contributing factor for the worse performance of RNAalifold is the integration of SHAPE data. There is a weakening of the information from experimental data with increasing number of homologs, because the pseudo energy from SHAPE reactivity is only applied to the free energy calculation of the particular sequence.

TurboFold II using SHAPE data on one or more sequences maintains a computation speed comparable to TurboFold II (with complexity  $O(H^2N^2 + HN^3)$  for  $H$  sequences of average length  $N$ ). The time performance on select sequence families is provided in Table S6.

## CONCLUSION

A new version of TurboFold II with the ability to include SHAPE mapping data for one or more of the RNA sequence homologs can substantially improve the structure prediction accuracies of the sequences that do not have SHAPE data. TurboFold II with the capability to include SHAPE mapping data for one or more sequences is available under the GNU license as part of the RNAstructure software package at: <http://rna.urmc.rochester.edu/RNAstructure.html>.

## SUPPORTING MATERIAL

Supporting Materials and Methods, fifteen figures, and six tables are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(17\)30689-6](http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30689-6).

## AUTHOR CONTRIBUTIONS

All authors planned experiments. Z.T. wrote code and performed experiments. Z.T. drafted the manuscript. All authors participated in the writing.

## ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) grants R01 GM097334 to G.S. and R01 GM076485 to D.H.M.

## REFERENCES

- Cech, T. R., and J. A. Steitz. 2014. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*. 157:77–94.
- Wu, L., and J. G. Belasco. 2008. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell*. 29:1–7.
- Doudna, J. A., and T. R. Cech. 2002. The chemical repertoire of natural ribozymes. *Nature*. 418:222–228.
- Gesteland, R. F., T. Cech, and J. F. Atkins. 2006. The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Seetin, M. G., and D. H. Mathews. 2012. RNA structure prediction: an overview of methods. *Methods Mol. Biol.* 905:99–122.
- Hofacker, I. L. 2014. Energy-directed RNA structure prediction. *Methods Mol. Biol.* 1097:71–84.
- Deigan, K. E., T. W. Li, ..., K. M. Weeks. 2009. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA*. 106:97–102.
- Quarrier, S., J. S. Martin, ..., A. Laederach. 2010. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*. 16:1108–1117.
- Washietl, S., I. L. Hofacker, ..., M. Kellis. 2012. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.* 40:4261–4272.
- Sloma, M. F., and D. H. Mathews. 2015. Improving RNA secondary structure prediction with structure mapping data. *Methods Enzymol.* 553:91–114.
- Mathews, D. H., M. D. Disney, ..., D. H. Turner. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*. 101:7287–7292.
- Hajdin, C. E., S. Bellaousov, ..., K. M. Weeks. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA*. 110:5498–5503.
- Mathews, D. H., J. Sabina, ..., D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.
- Eddy, S. R. 2014. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.* 43:433–456.
- Zarringhalam, K., M. M. Meyer, ..., P. Clote. 2012. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS One*. 7:e45160.
- Ouyang, Z., M. P. Snyder, and H. Y. Chang. 2013. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* 23:377–387.
- Deng, F., M. Ledda, ..., S. Aviran. 2016. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*. 22:1109–1119.
- McGinnis, J. L., J. A. Dunkle, ..., K. M. Weeks. 2012. The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.* 134:6617–6624.
- Merino, E. J., K. A. Wilkinson, ..., K. M. Weeks. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* 127:4223–4231.
- Sükösd, Z., M. S. Swenson, ..., C. E. Heitsch. 2013. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Res.* 41:2807–2816.
- Kertesz, M., Y. Wan, ..., E. Segal. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature*. 467:103–107.



22. Underwood, J. G., A. V. Uzilov, ..., D. Haussler. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*. 7:995–1001.
23. Talkish, J., G. May, ..., C. J. McManus. 2014. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*. 20:713–720.
24. Ding, Y., Y. Tang, ..., S. M. Assmann. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*. 505:696–700.
25. Spitale, R. C., P. Crisalli, ..., H. Y. Chang. 2013. RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* 9:18–20.
26. Rouskin, S., M. Zubradt, ..., J. S. Weissman. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*. 505:701–705.
27. Cordero, P., and R. Das. 2015. Rich RNA structure landscapes revealed by mutate-and-map analysis. *PLOS Comput. Biol.* 11:e1004473.
28. Puto, T., L. P. Kozłowski, ..., J. M. Bujnicki. 2014. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.* 42:5403–5406.
29. Havgaard, J. H., and J. Gorodkin. 2014. RNA structural alignments, part I: Sankoff-based approaches for structural alignments. *Methods Mol. Biol.* 1097:275–290.
30. Asai, K., and M. Hamada. 2014. RNA structural alignments, part II: non-Sankoff approaches for structural alignments. *Methods Mol. Biol.* 1097:291–301.
31. Harmanci, A. O., G. Sharma, and D. H. Mathews. 2011. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*. 12:108.
32. Harmanci, A. O., G. Sharma, and D. H. Mathews. 2007. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*. 8:130.
33. Hofacker, I. L., S. H. Bernhart, and P. F. Stadler. 2004. Alignment of RNA base pairing probability matrices. *Bioinformatics*. 20:2222–2227.
34. Knudsen, B., and J. Hein. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 31:3423–3428.
35. Lu, Z. J., J. W. Gloor, and D. H. Mathews. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*. 15:1805–1813.
36. Do, C. B., M. S. Mahabhashyam, ..., S. Batzoglou. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15:330–340.
37. Lavender, C. A., R. Lorenz, ..., K. M. Weeks. 2015. Model-Free RNA sequence and structure alignment informed by SHAPE probing reveals a conserved alternate secondary structure for 16S rRNA. *PLOS Comput. Biol.* 11:e1004126.
38. Bernhart, S. H., I. L. Hofacker, ..., P. F. Stadler. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*. 9:474.
39. Ding, Y., and C. E. Lawrence. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* 31:7280–7301.
40. Mathews, D. H. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*. 10:1178–1190.
41. Turner, D. H., and D. H. Mathews. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* 38:D280–D282.
42. Durbin, R., S. R. Eddy, ..., G. Mitchison. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, United Kingdom.
43. Do, C. B., D. A. Woods, and S. Batzoglou. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. 22:e90–e98.
44. Bellaousov, S., and D. H. Mathews. 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*. 16:1870–1880.
45. Nawrocki, E. P., S. W. Burge, ..., R. D. Finn. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43:D130–D137.
46. Fu, Y., G. Sharma, and D. H. Mathews. 2014. Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res.* 42:13939–13948.
47. Gutell, R. R., J. C. Lee, and J. J. Cannone. 2002. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12:301–310.
48. Woodson, S. A., and D. M. Crothers. 1987. Proton nuclear magnetic resonance studies on bulge-containing DNA oligonucleotides from a mutational hot-spot sequence. *Biochemistry*. 26:904–912.
49. Znosko, B. M., S. B. Silvestri, ..., M. J. Serra. 2002. Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*. 41:10406–10417.
50. R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
51. Xu, Z., A. Almudevar, and D. H. Mathews. 2012. Statistical evaluation of improvement in RNA secondary structure prediction. *Nucleic Acids Res.* 40:e26.
52. Lorenz, R., S. H. Bernhart, ..., I. L. Hofacker. 2011. ViennaRNA package 2.0. *Algorithms Mol. Biol.* 6:26.
53. Larkin, M. A., G. Blackshields, ..., D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*. 23:2947–2948.
54. Mauger, D. M., N. A. Siegfried, and K. M. Weeks. 2013. The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Lett.* 587:1180–1188.
55. Reuter, J. S., and D. H. Mathews. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*. 11:129.

**Biophysical Journal, Volume 113**

**Supplemental Information**

**Modeling RNA Secondary Structure with Sequence Comparison and  
Experimental Mapping Data**

**Zhen Tan, Gaurav Sharma, and David H. Mathews**

## **Supplementary information for**

### **“Modeling RNA secondary structure with sequence comparison and experimental mapping data”**

**Z. Tan, G. Sharma, and D. H. Mathews**

Details are provided for dataset used in benchmarking (Section 1), structure modeling accuracy (Section 2), parameter optimization methods (Section 2), sequences used in parameter optimization, software efficiency test (Section 3), and benchmarking (Section 4).

# Section 1. Dataset information:

Family	<i>H</i>	Average sequence length	Standard deviation	Average MEA sensitivity	Standard deviation	Average MEA PPV	Standard deviation
tRNA	5 sequences	75.7	3.5	0.76	0.23	0.75	0.24
	10 sequences	76.2	4.7	0.77	0.23	0.74	0.25
	20 sequences	76.3	4.8	0.77	0.21	0.74	0.23
cGMP riboswitch	5 sequences	89.0	8.3	0.86	0.19	0.33	0.12
	10 sequences	87.9	6.9	0.81	0.26	0.31	0.13
	20 sequences	87.5	6.5	0.81	0.25	0.31	0.13
TPP riboswitch	5 sequences	101.5	16.8	0.54	0.29	0.43	0.28
	10 sequences	104.4	13.9	0.55	0.29	0.44	0.27
	20 sequences	106.1	13.1	0.55	0.29	0.43	0.27
SAM I riboswitch	5 sequences	111.3	13.9	0.83	0.18	0.68	0.17
	10 sequences	111.9	14.1	0.82	0.17	0.67	0.16
	20 sequences	111.9	15.3	0.84	0.16	0.68	0.15
5S rRNA	5 sequences	117.7	4.6	0.64	0.24	0.62	0.24
	10 sequences	117.8	3.2	0.56	0.25	0.55	0.25
	20 sequences	117.8	4.2	0.57	0.27	0.54	0.26
M-box riboswitch	5 sequences	164.7	8.5	0.64	0.15	0.61	0.15
	10 sequences	167.1	8.5	0.66	0.17	0.62	0.16
	20 sequences	167.8	7.3	0.66	0.15	0.63	0.14
lysine riboswitch	5 sequences	179.1	6.8	0.76	0.17	0.71	0.15
	10 sequences	183.5	12.6	0.65	0.22	0.60	0.20
	20 sequences	182.7	10.7	0.68	0.22	0.63	0.21
HCV	5 sequences	267.4	66.1	0.50	0.16	0.46	0.16
	10 sequences	250.7	62.9	0.47	0.17	0.43	0.17
	20 sequences	251.0	60.5	0.48	0.18	0.43	0.17
Group I intron	5 sequences	431.1	51.0	0.61	0.16	0.58	0.15
	10 sequences	433.3	52.7	0.60	0.16	0.59	0.16
	20 sequences	433.8	54.0	0.61	0.16	0.59	0.16
23S rRNA	5 sequences	2919.4	51.8	0.52	0.53	0.08	0.07
	10 sequences	2928.8	62.6	0.51	0.52	0.02	0.04
	20 sequences	2924.3	56.4	0.52	0.51	0.01	0.06

Table S1. **Summary statistics on the sets of sequences selected for testing.** Mean and standard deviation of sequence length, sensitivity and PPV of MEA structure prediction are shown for sequences from each RNA family in the test sets of homologs used.



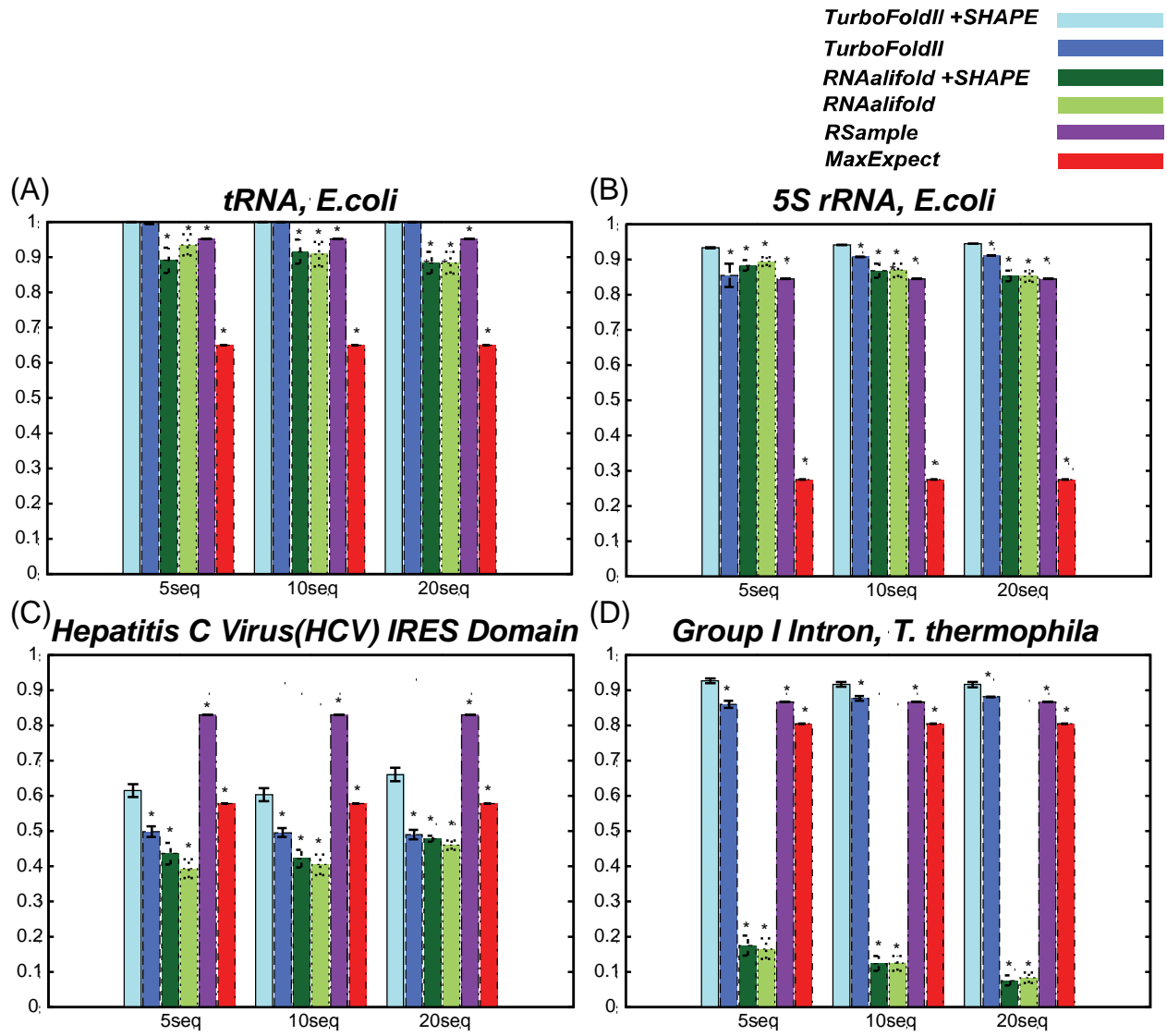
Family	Total number of distinct sequences	Total number of sequences in database
tRNA	627	9245
cGMP riboswitch	150	155
TPP riboswitch	97	109
SAM I riboswitch	272	433
5S rRNA	429	710
M-box riboswitch	138	157
Lysine riboswitch	45	47
HCV	74	79
Group I intron	437	816
23S rRNA	35	35

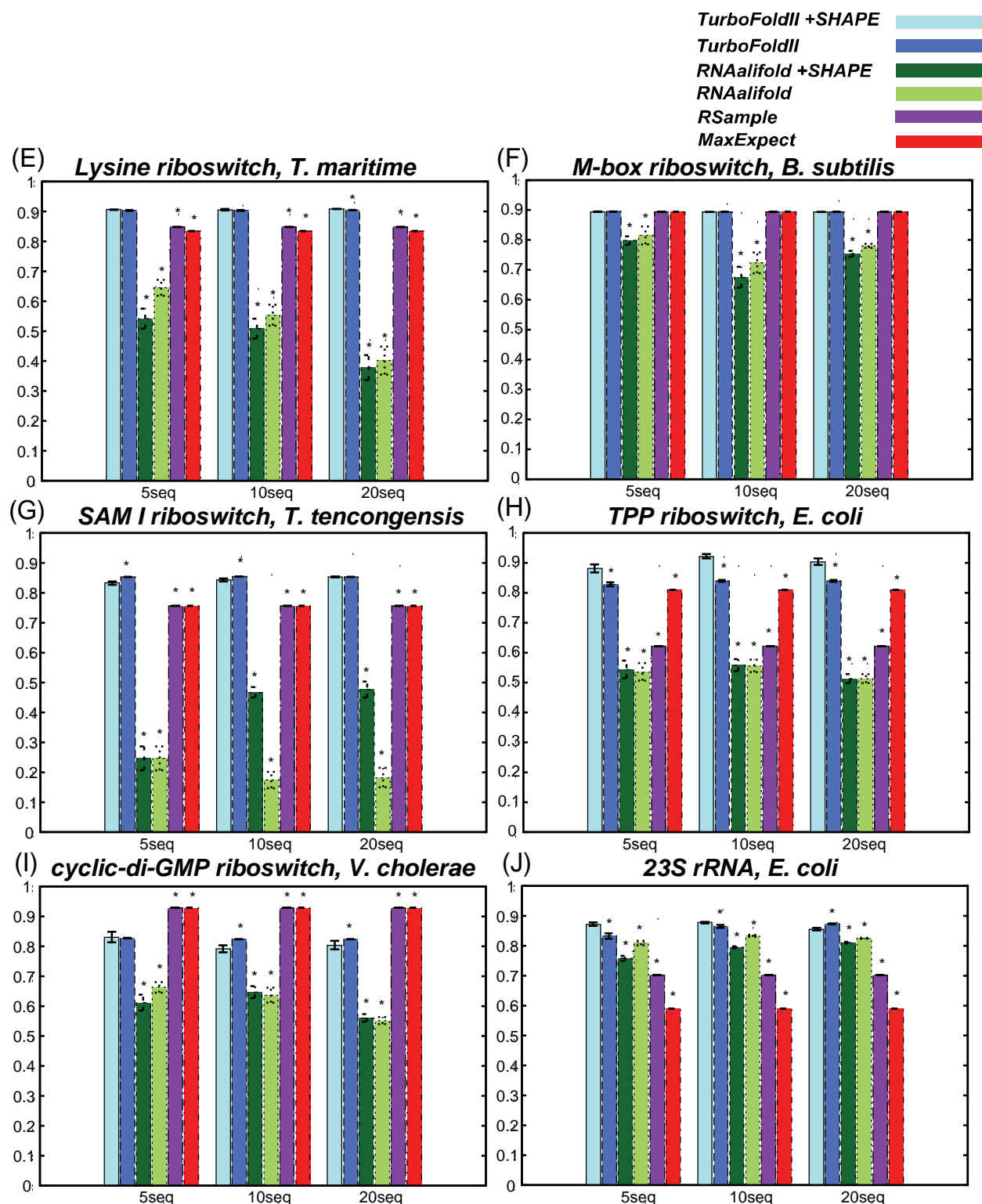
Table S2. **Number of distinct sequences on the sets of sequences selected for testing.** Number of distinct sequences from each RNA family in test sets and the total number of sequences available in database are shown.

Family	Sequence with SHAPE reactivity data
tRNA	<i>E. coli</i>
cGMP riboswitch	<i>V. cholerae</i>
TPP riboswitch	<i>E. coli</i>
SAM I riboswitch	<i>T. tencongensis</i>
5S rRNA	<i>E. coli</i>
M-box riboswitch	<i>B. subtilis</i>
Lysine riboswitch	<i>T. maritime</i>
HCV	<i>Hepatitis C virus IRES domain</i>
Group I intron	<i>T. thermophila</i>
23S rRNA	<i>E. coli</i>

Table S3. **List of sequences with SHAPE reactivity data for each family.**

## Section 2. Structure prediction accuracy:





**Figure S1. Average F1 score of structure predictions of sequences that did not have SHAPE mapping data.** F1 score of structures predictions obtained by running the methods with 5, 10, or 20 input sequences on (A) tRNA, (B) 5S rRNA, (C) hepatitis C virus IRES domain, (D) group I intron, (E) lysine riboswitch, (F) M-box riboswitch, (G) SAM I riboswitch, (H) TPP riboswitch,

(I) cyclic-di-GMP riboswitch, and (J) 23S rRNA test datasets. The star (\*) above the bar for a method indicates that the difference in F1 score between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



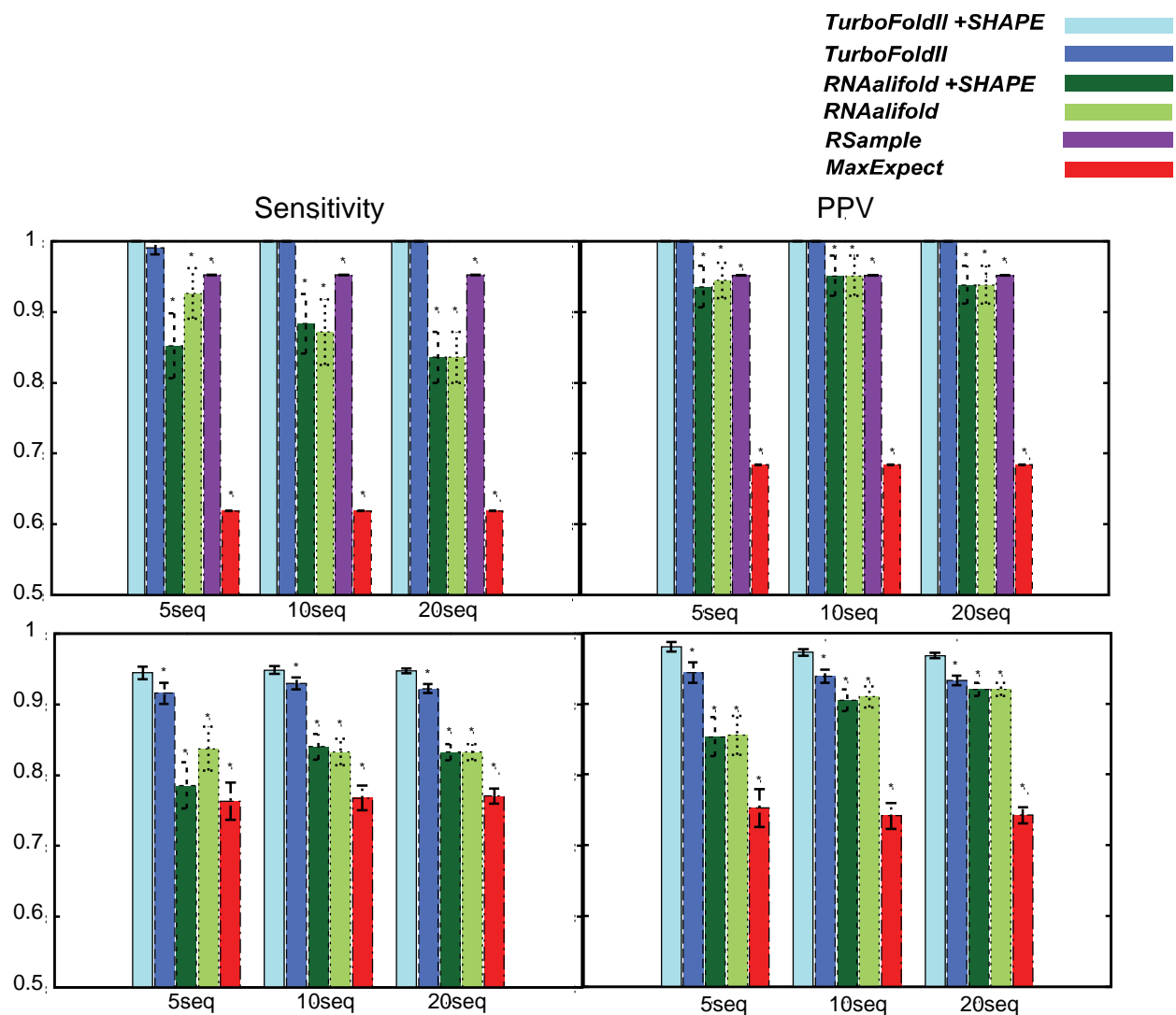


Figure S2. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on tRNA test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on tRNA test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

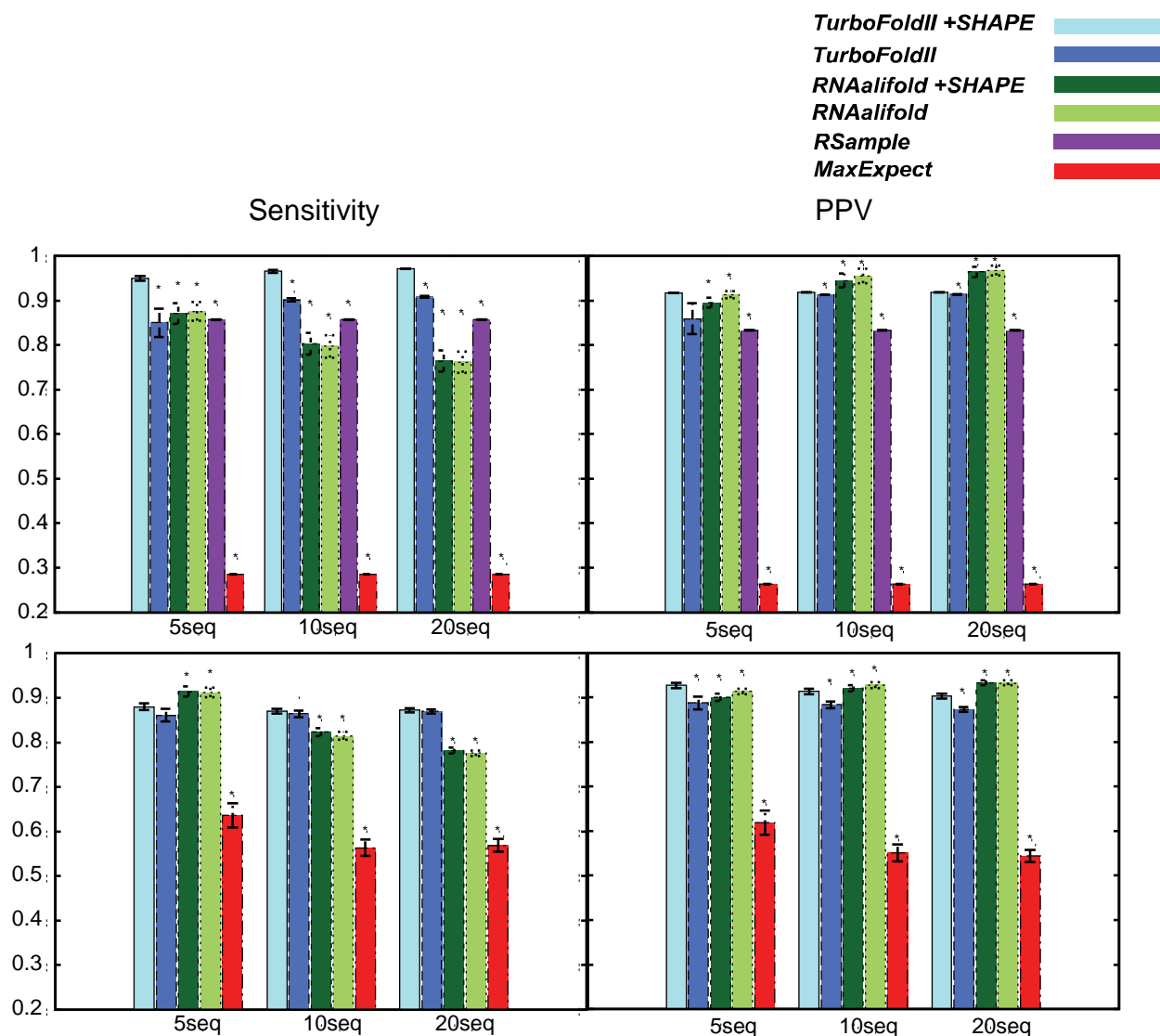


Figure S3. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on 5S rRNA test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on 5S rRNA test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

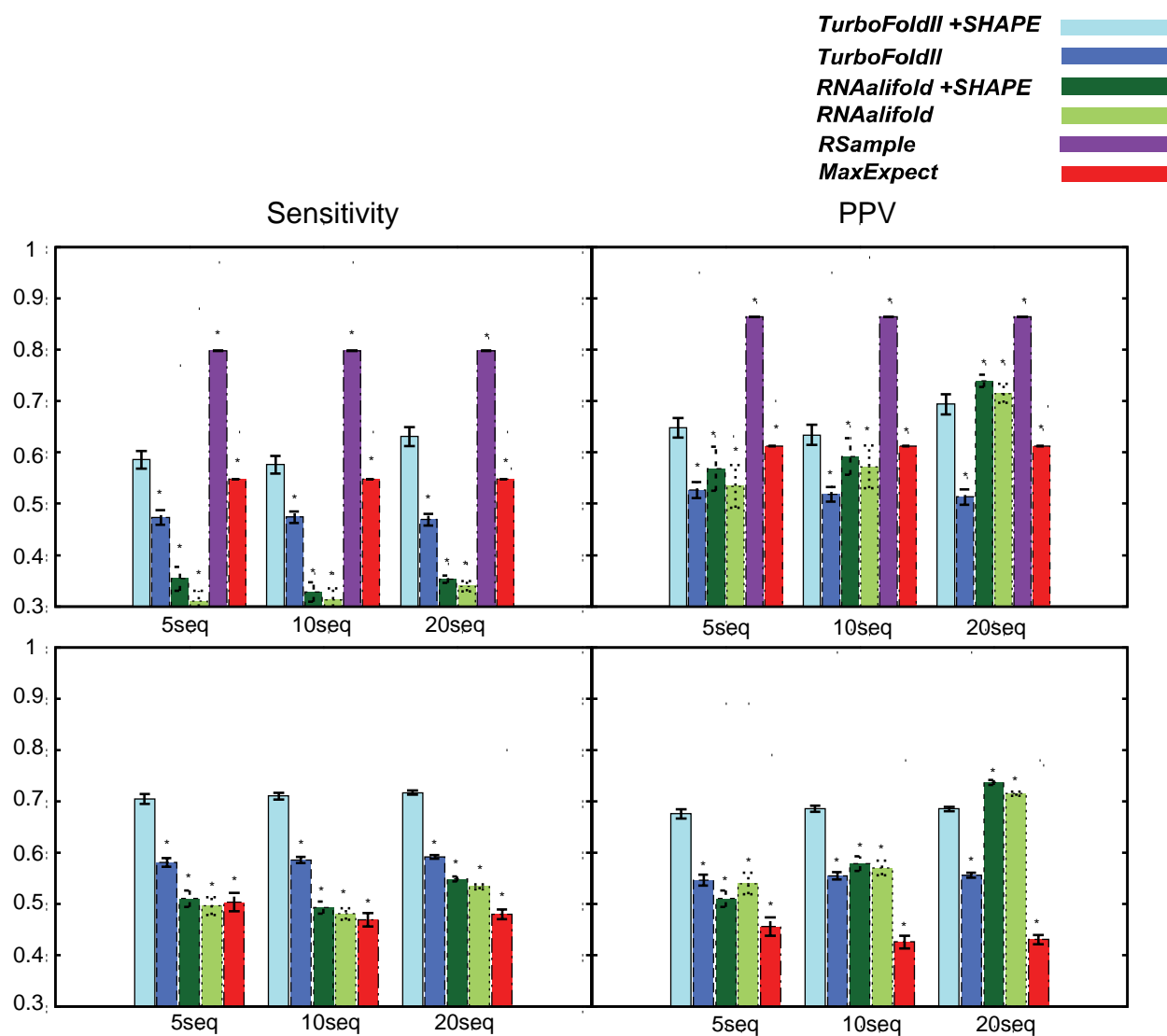


Figure S4. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on hepatitis C virus (HCV) IRES domain test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with 5, 10, or 20 input sequences on hepatitis C virus (HCV) IRES domain test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

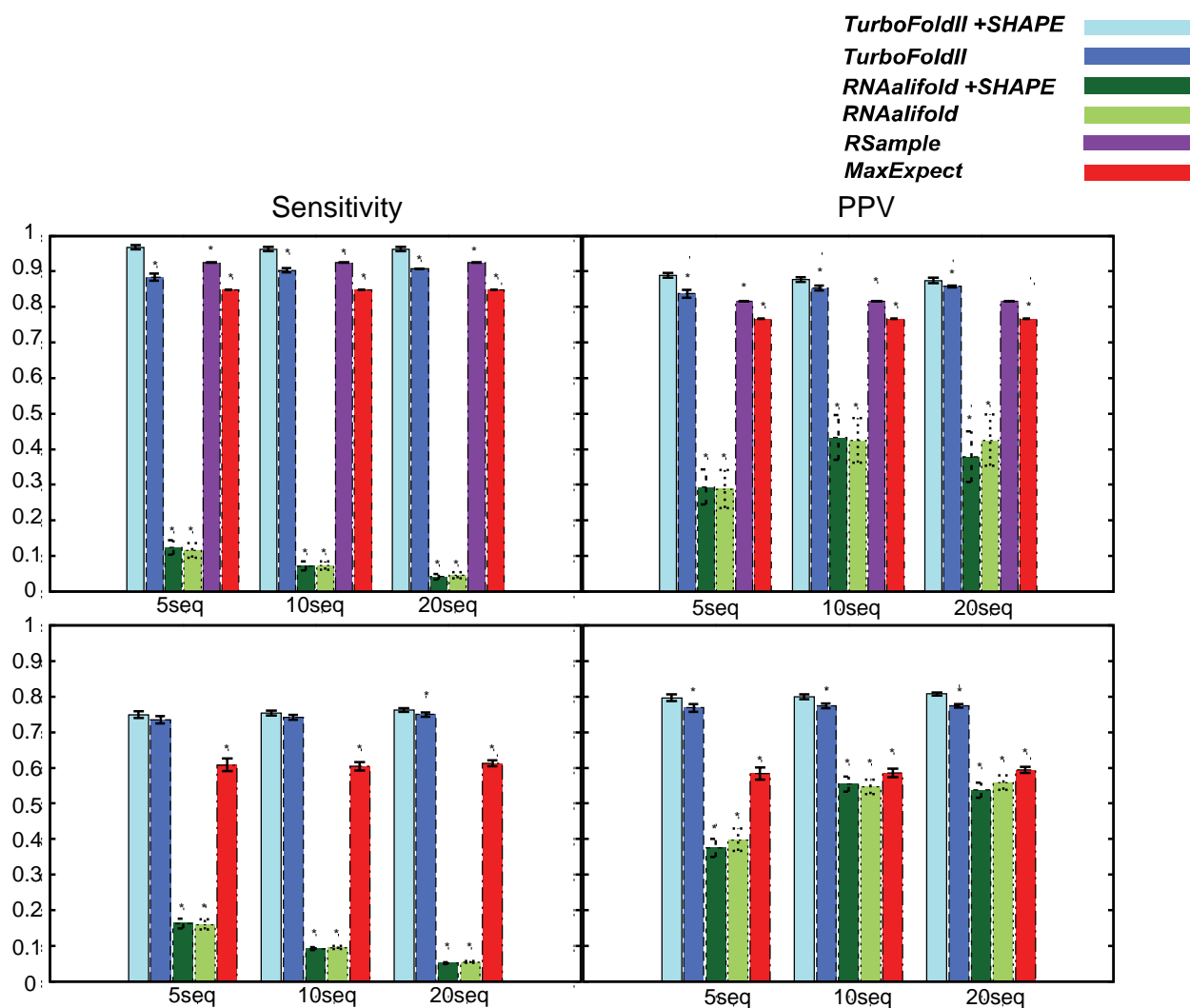


Figure S5. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on group I intron test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on group I intron test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



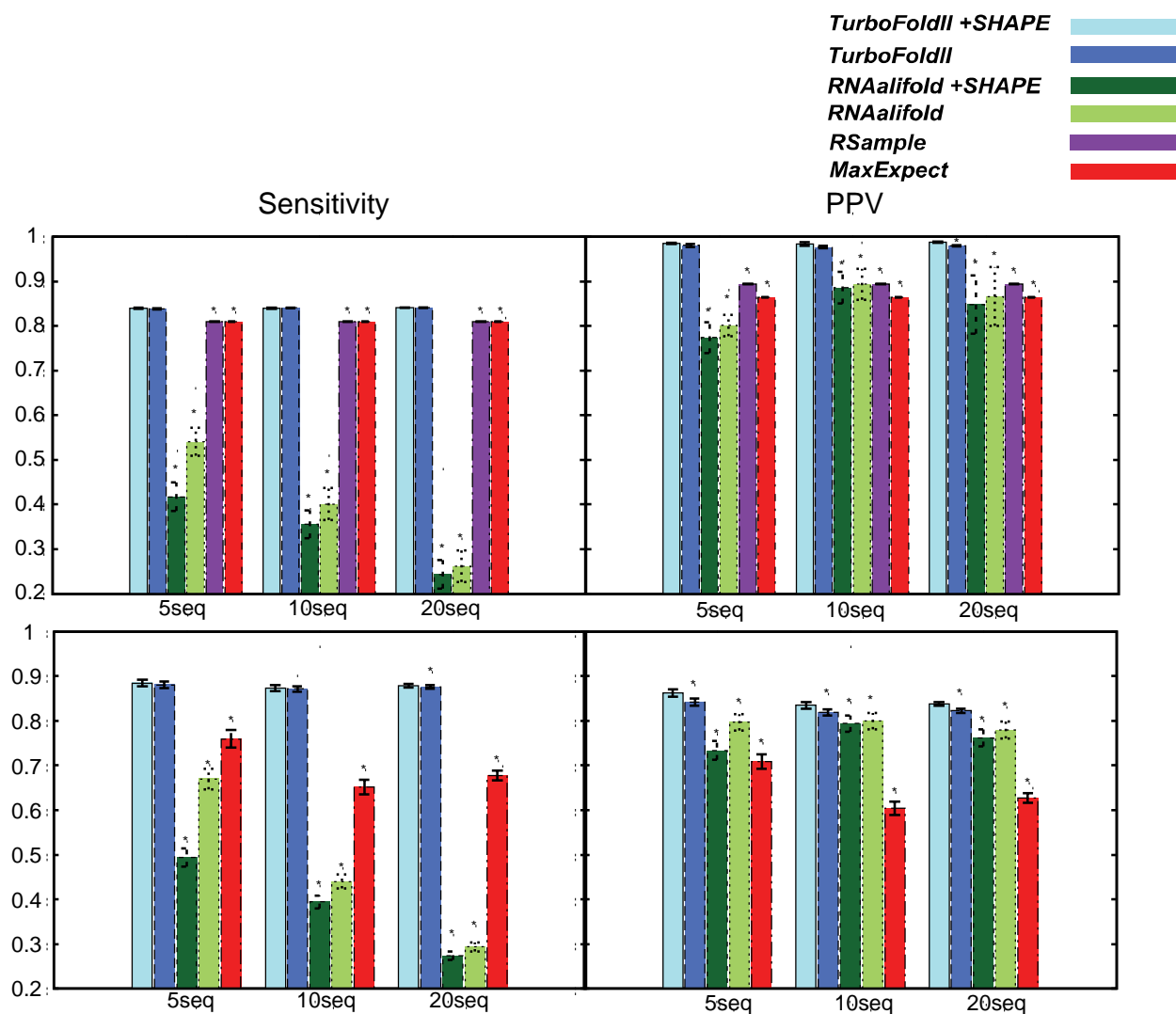
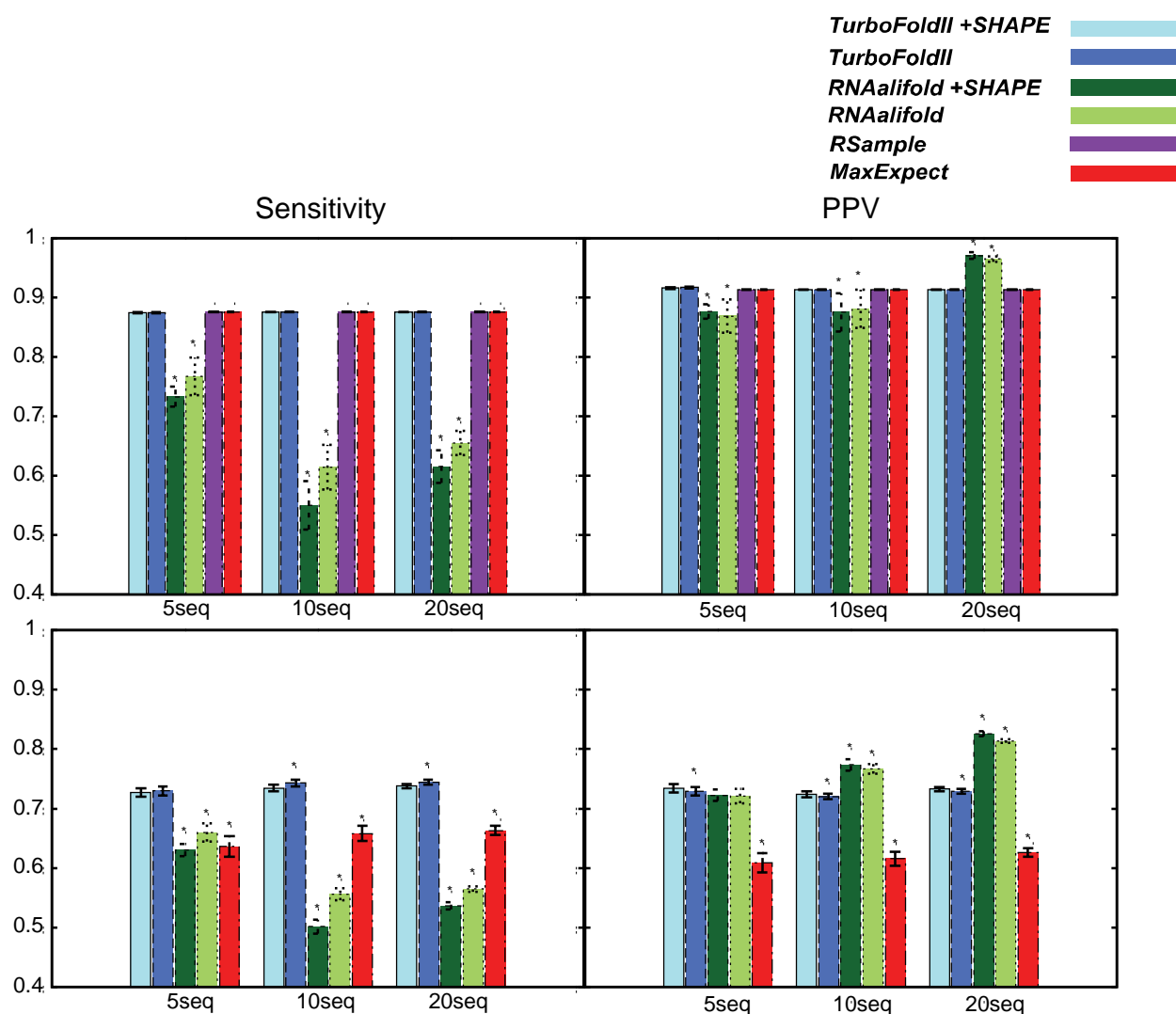
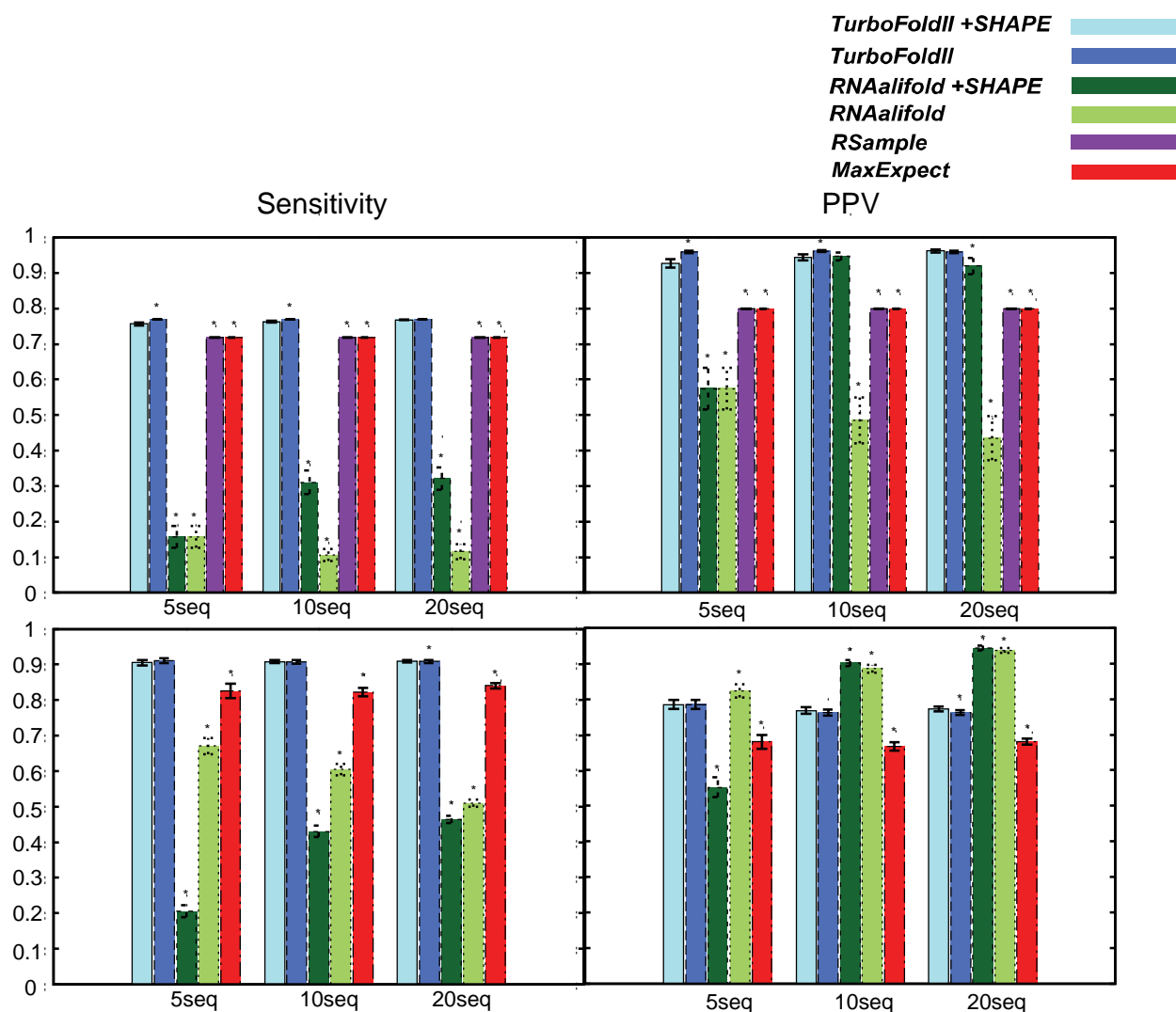


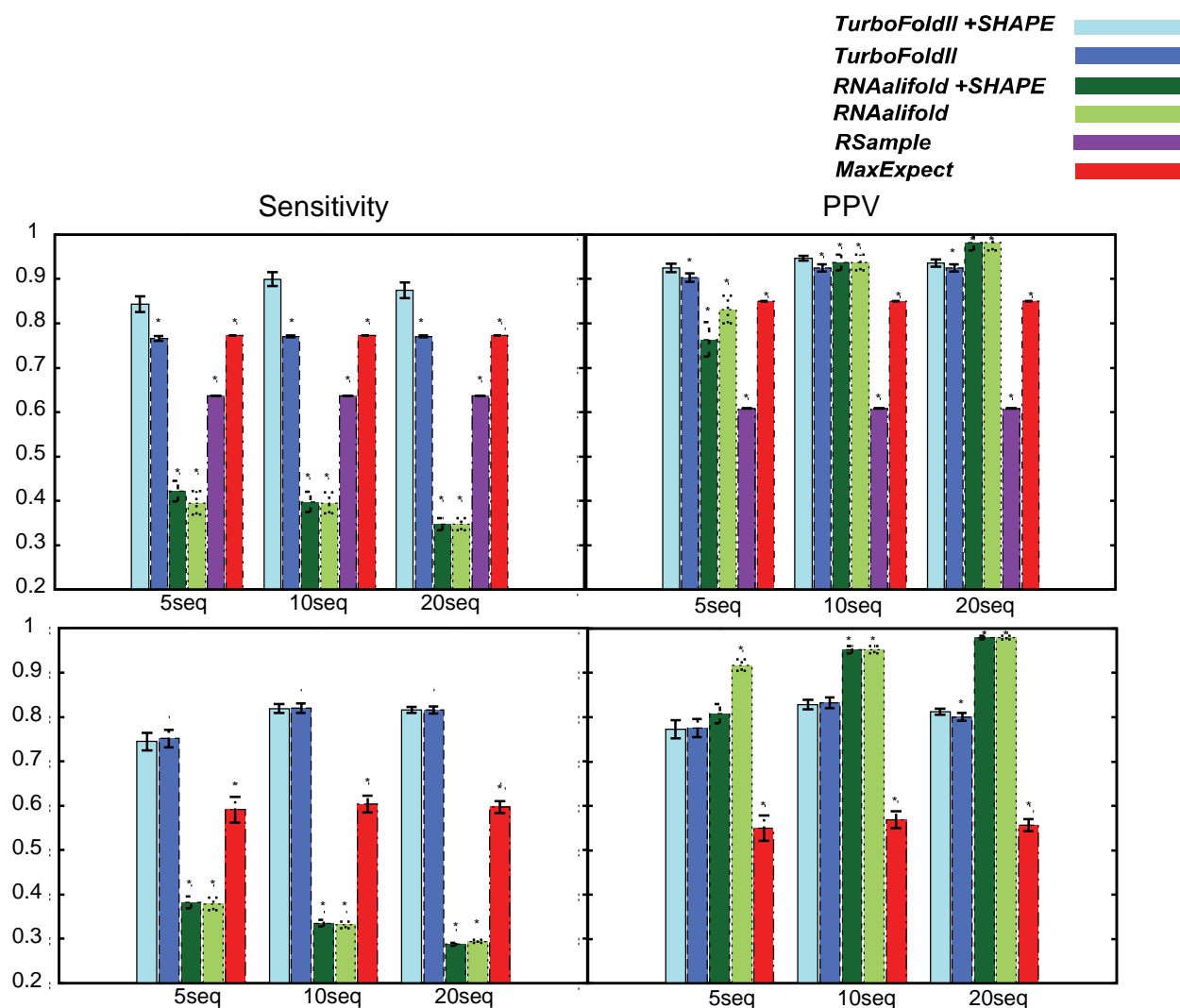
Figure S6. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on lysine riboswitch test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with 5, 10, or 20 input sequences on lysine riboswitch test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



**Figure S7. Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on M-box riboswitch test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with 5, 10, or 20 input sequences on M-box riboswitch test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



**Figure S8. Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on SAM I riboswitch test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on SAM I riboswitch test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



**Figure S9. Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on TPP riboswitch test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on TPP riboswitch test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

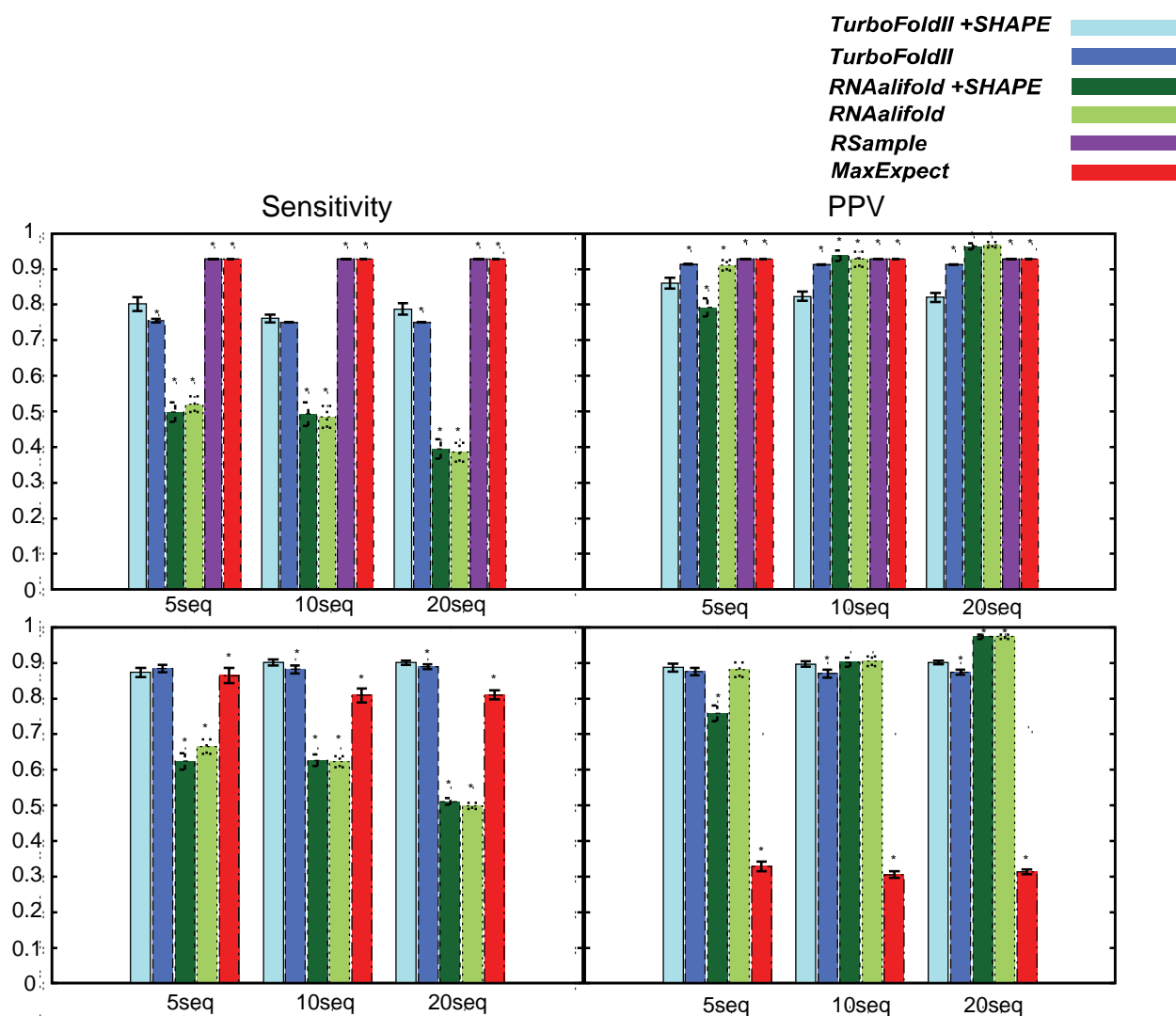


Figure S10. Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on cyclic-di-GMP riboswitch test datasets. Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on cyclic-di-GMP riboswitch test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

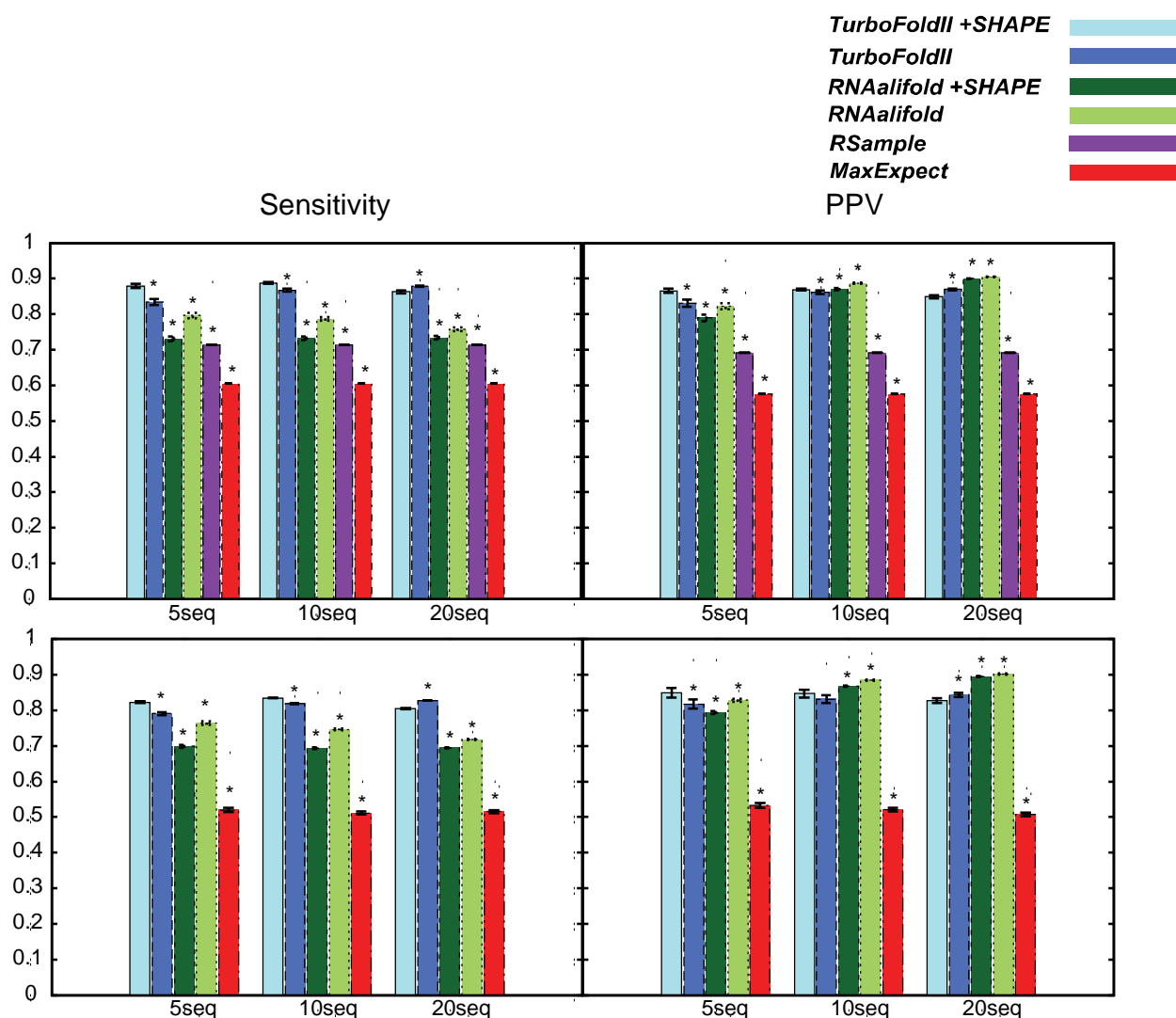


Figure S11. **Average Sensitivity and PPV of structure predictions of sequences that have SHAPE mapping data (top) and sequences that do not have SHAPE mapping data (bottom) on 23S rRNA test datasets.** Sensitivity and PPV of structures predictions obtained by running the methods with  $H = 5, 10$ , or  $20$  input sequences on 23S rRNA test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity or PPV between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.

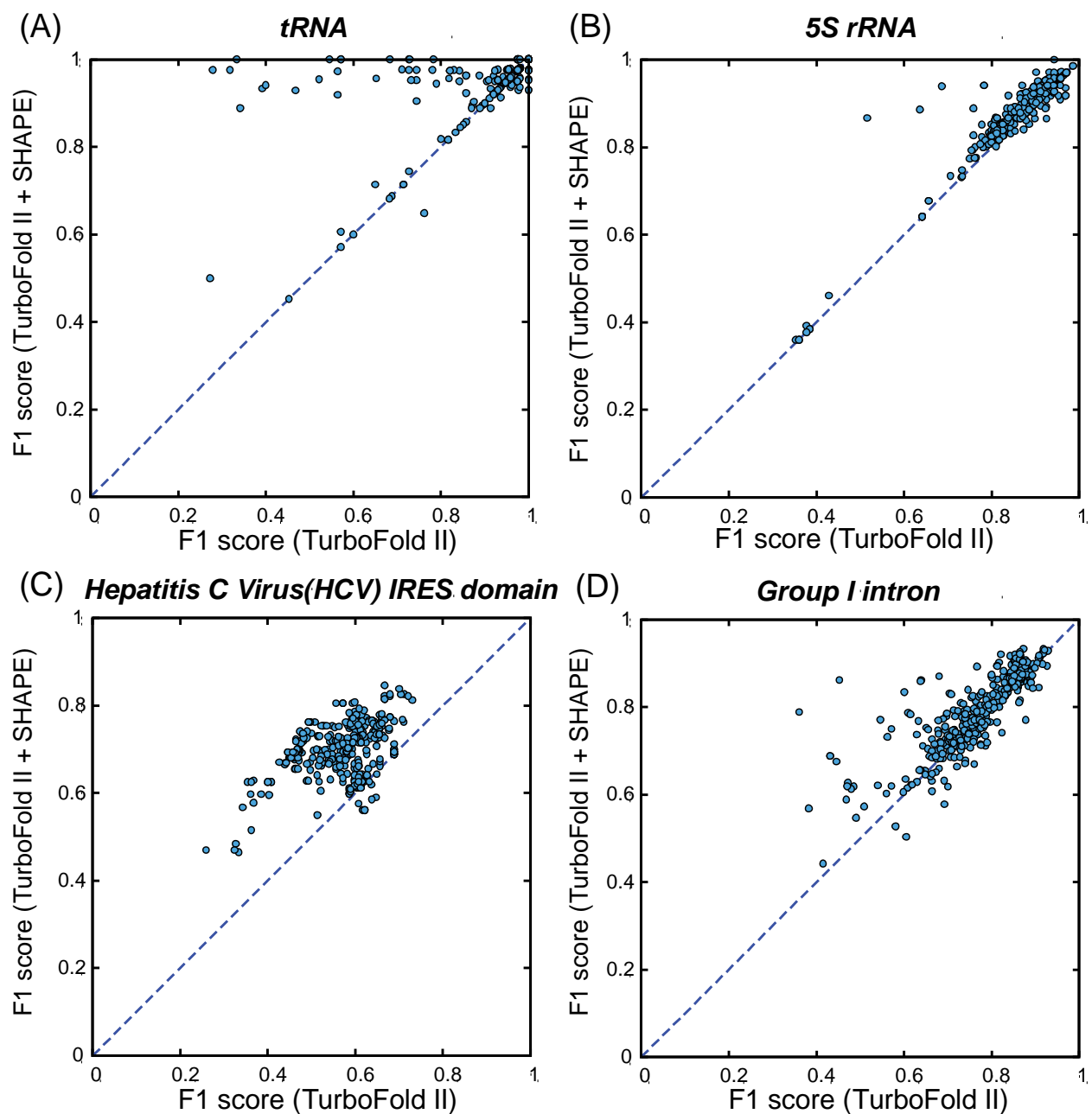


Figure S12. **Scatter plots of F1 score of structure predictions obtained with TurboFold II and TurboFold II + SHAPE for sequences (20 sequence groups) that did not have SHAPE mapping data.** The F1 scores of structures predictions are obtained by running the methods with  $H = 20$  input sequences on tRNA, 5S rRNA, hepatitis C virus IRES domain, and group I intron RNA test datasets. Each point represents the F1 scores by TurboFold II and TurboFold II + SHAPE for one sequence.

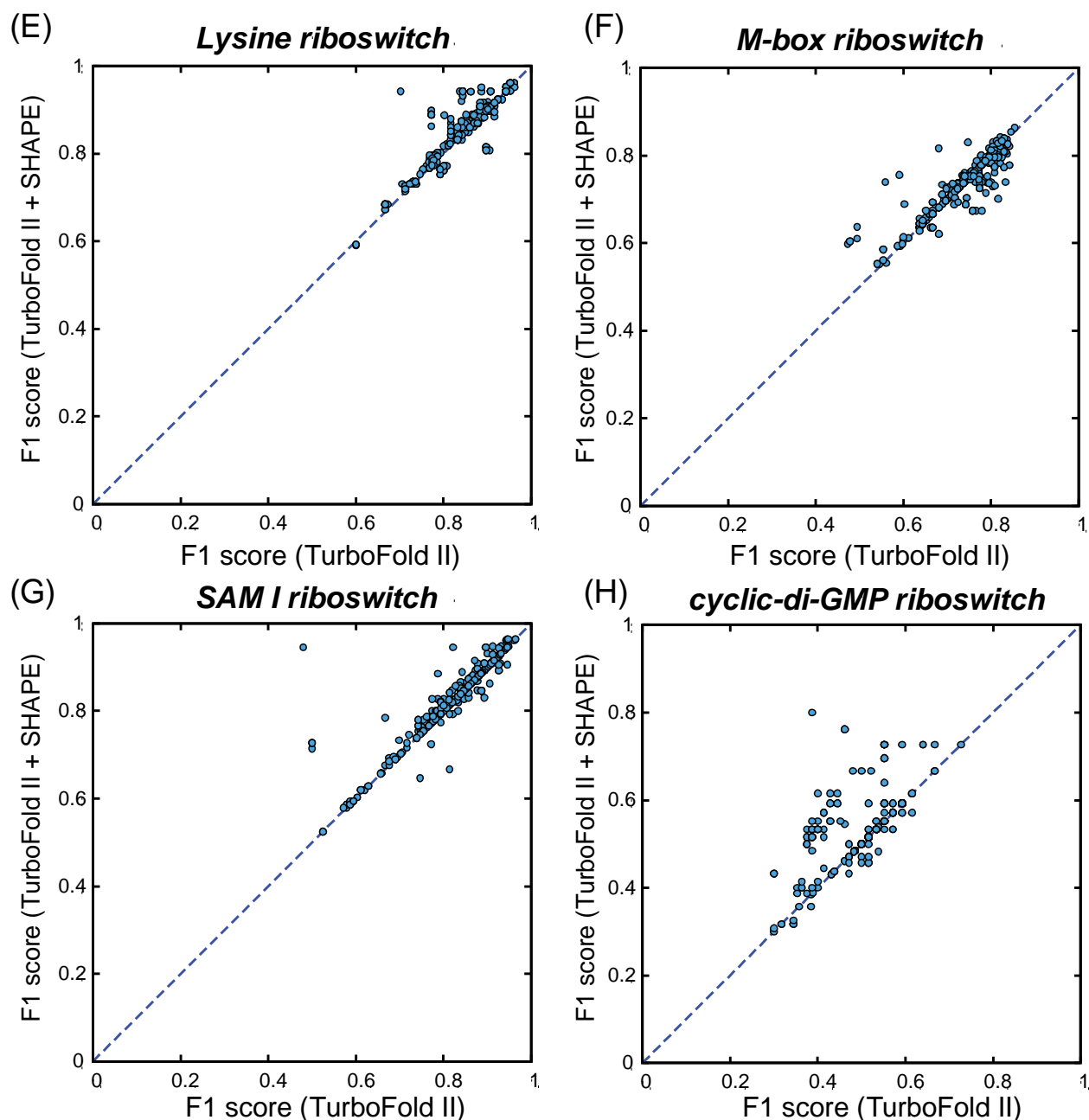


Figure S12. **Scatter plots of F1 score of structure predictions obtained with TurboFold II and TurboFold II+SHAPE for sequences (20 sequence groups) that do not have SHAPE mapping data.** F1 score of structures predictions obtained by running the methods with  $H = 20$  input sequences on lysine riboswitch, M-box riboswitch, SAM I riboswitch, and cyclic-di-GMP riboswitch test datasets. Each dot represents the F1 scores by TurboFold II and TurboFold II+SHAPE.



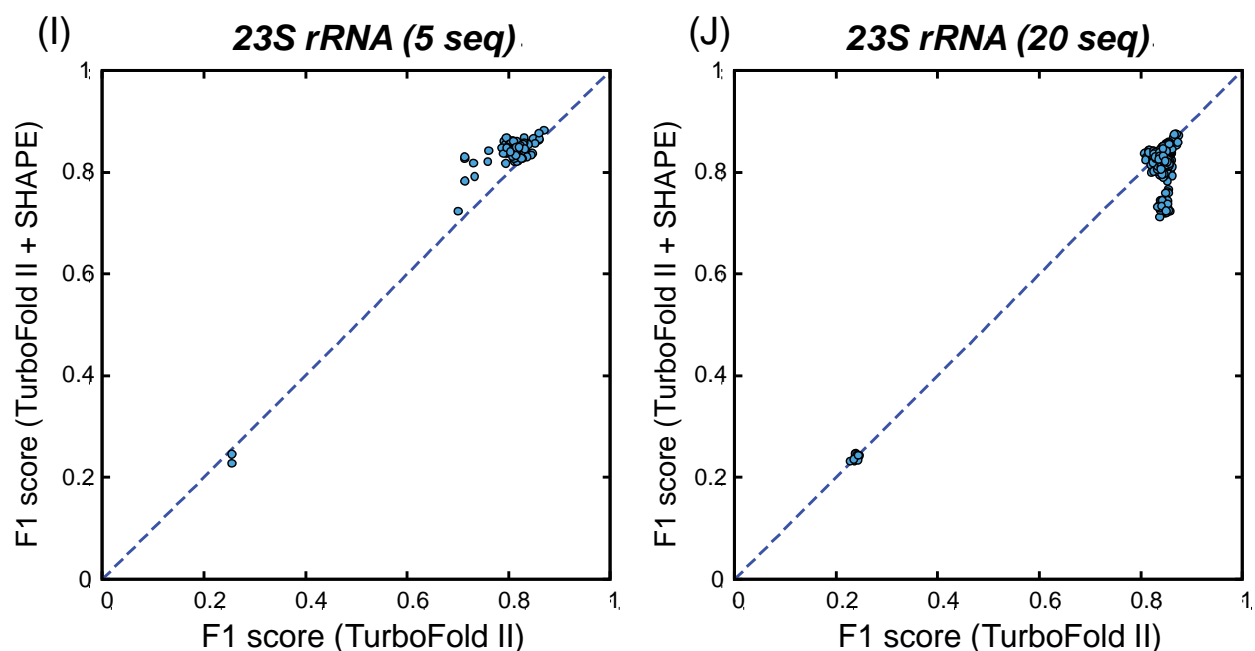


Figure S12. **Scatter plots of F1 score of structure predictions obtained with TurboFold II and TurboFold II+SHAPE for sequences (20 sequence groups) that do not have SHAPE mapping data.** F1 score of structures predictions obtained by running the methods with 5 input sequences (left) and  $H = 20$  input sequences (right) on (A) tRNA, (B) 5S rRNA, (C) hepatitis C virus IRES domain, (D) group I intron, (E) lysine riboswitch, (F) M-box riboswitch, (G) SAM I riboswitch, (H) cyclic-di-GMP riboswitch, (I) 23S rRNA (5 sequences), and (J) 23S rRNA (20 sequences) test datasets. Each dot represents the F1 scores by TurboFold II and TurboFold II + SHAPE.

Table S4. Average structure prediction sensitivity and PPV on sequences without SHAPE data for each method on each dataset:

<b>5S rRNA</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.880	0.927	0.871	0.913	0.873	0.903
<b>TurboFold II</b>	0.861	0.888	0.864	0.883	0.869	0.873
<b>RNAalifold + SHAPE</b>	0.914	0.900	0.823	0.921	0.782	0.932
<b>RNAalifold</b>	0.912	0.914	0.815	0.928	0.776	0.932
<b>MaxExpect</b>	0.636	0.619	0.564	0.551	0.569	0.544

<b>Group I intron</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.749	0.797	0.754	0.800	0.763	0.807
<b>TurboFold II</b>	0.735	0.769	0.742	0.774	0.750	0.775
<b>RNAalifold + SHAPE</b>	0.163	0.375	0.092	0.554	0.052	0.537
<b>RNAalifold</b>	0.160	0.398	0.095	0.547	0.054	0.558
<b>MaxExpect</b>	0.608	0.584	0.604	0.585	0.612	0.594

<b>HCV</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.705	0.676	0.710	0.686	0.717	0.685
<b>TurboFold II</b>	0.581	0.547	0.586	0.555	0.592	0.557
<b>RNAalifold + SHAPE</b>	0.510	0.510	0.493	0.579	0.549	0.737
<b>RNAalifold</b>	0.496	0.540	0.481	0.570	0.534	0.715
<b>MaxExpect</b>	0.504	0.456	0.469	0.426	0.480	0.431

tRNA						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.945	0.981	0.949	0.973	0.948	0.968
<b>TurboFold II</b>	0.916	0.944	0.930	0.939	0.922	0.933
<b>RNAalifold + SHAPE</b>	0.786	0.853	0.840	0.905	0.833	0.920
<b>RNAalifold</b>	0.837	0.856	0.833	0.910	0.833	0.920
<b>MaxExpect</b>	0.763	0.752	0.768	0.742	0.771	0.742

TPP riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.744	0.773	0.819	0.829	0.816	0.812
<b>TurboFold II</b>	0.752	0.775	0.820	0.833	0.816	0.801
<b>RNAalifold + SHAPE</b>	0.382	0.808	0.335	0.952	0.288	0.980
<b>RNAalifold</b>	0.379	0.917	0.332	0.953	0.294	0.980
<b>MaxExpect</b>	0.535	0.428	0.547	0.436	0.552	0.431

SAM I riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.905	0.784	0.908	0.768	0.910	0.772
<b>TurboFold II</b>	0.911	0.785	0.908	0.762	0.908	0.762
<b>RNAalifold + SHAPE</b>	0.206	0.552	0.430	0.902	0.464	0.945
<b>RNAalifold</b>	0.671	0.824	0.604	0.886	0.510	0.937
<b>MaxExpect</b>	0.826	0.680	0.822	0.667	0.840	0.681

M-box riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.727	0.734	0.734	0.724	0.738	0.733
<b>TurboFold II</b>	0.730	0.729	0.743	0.720	0.744	0.729
<b>RNAalifold + SHAPE</b>	0.630	0.722	0.502	0.774	0.536	0.826
<b>RNAalifold</b>	0.660	0.721	0.556	0.767	0.565	0.814
<b>MaxExpect</b>	0.636	0.608	0.658	0.615	0.663	0.626

Lysine riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.885	0.862	0.873	0.834	0.878	0.838
<b>TurboFold II</b>	0.880	0.842	0.871	0.819	0.875	0.823
<b>RNAalifold + SHAPE</b>	0.494	0.733	0.394	0.794	0.274	0.762
<b>RNAalifold</b>	0.670	0.796	0.440	0.799	0.294	0.779
<b>MaxExpect</b>	0.760	0.709	0.651	0.604	0.677	0.627

Cyclic-di-GMP riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.874	0.887	0.902	0.897	0.900	0.901
<b>TurboFold II</b>	0.884	0.876	0.882	0.871	0.889	0.874
<b>RNAalifold + SHAPE</b>	0.624	0.759	0.626	0.902	0.511	0.974
<b>RNAalifold</b>	0.665	0.881	0.623	0.904	0.498	0.974
<b>MaxExpect</b>	0.865	0.329	0.809	0.306	0.810	0.313

<b>23S rRNA</b>						
Prediction Method	<i>H</i> = 5 sequences		<i>H</i> = 10 sequences		<i>H</i> = 20 sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.823	0.868	0.834	0.876	0.803	0.848
<b>TurboFold II</b>	0.788	0.834	0.817	0.858	0.826	0.865
<b>RNAalifold + SHAPE</b>	0.699	0.793	0.693	0.867	0.696	0.895
<b>RNAalifold</b>	0.764	0.828	0.746	0.885	0.718	0.902
<b>MaxExpect</b>	0.520	0.533	0.511	0.521	0.515	0.507

Table S5. Average structure prediction sensitivity and PPV on sequences with SHAPE data for each method on each dataset:

<b>5S rRNA</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.950	0.917	0.966	0.918	0.971	0.919
<b>TurboFold II</b>	0.850	0.859	0.901	0.913	0.909	0.914
<b>RNAalifold + SHAPE</b>	0.871	0.896	0.803	0.945	0.764	0.964
<b>RNAalifold</b>	0.876	0.914	0.797	0.955	0.761	0.967
<b>Rsample</b>	0.857	0.833	0.857	0.833	0.857	0.833
<b>MaxExpect</b>	0.286	0.263	0.286	0.263	0.286	0.263
<b>Group I intron</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.968	0.889	0.962	0.877	0.963	0.874
<b>TurboFold II</b>	0.884	0.837	0.903	0.853	0.907	0.858
<b>RNAalifold + SHAPE</b>	0.124	0.294	0.072	0.433	0.042	0.379
<b>RNAalifold</b>	0.116	0.288	0.073	0.425	0.046	0.425
<b>Rsample</b>	0.924	0.816	0.924	0.816	0.924	0.816
<b>MaxExpect</b>	0.849	0.766	0.849	0.766	0.849	0.766
<b>HCV</b>						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.586	0.648	0.576	0.634	0.631	0.694
<b>TurboFold II</b>	0.473	0.527	0.474	0.519	0.469	0.513
<b>RNAalifold + SHAPE</b>	0.354	0.568	0.328	0.592	0.353	0.740
<b>RNAalifold</b>	0.311	0.534	0.313	0.572	0.339	0.715
<b>Rsample</b>	0.798	0.864	0.798	0.864	0.798	0.864
<b>MaxExpect</b>	0.548	0.612	0.548	0.612	0.548	0.612

tRNA						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	1.000	1.000	1.000	1.000	1.000	1.000
<b>TurboFold II</b>	0.990	1.000	1.000	1.000	1.000	1.000
<b>RNAalifold + SHAPE</b>	0.852	0.936	0.883	0.951	0.836	0.938
<b>RNAalifold</b>	0.926	0.944	0.871	0.951	0.836	0.938
<b>Rsample</b>	0.952	0.952	0.952	0.952	0.952	0.952
<b>MaxExpect</b>	0.619	0.684	0.619	0.684	0.619	0.684

TPP riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.843	0.925	0.900	0.947	0.875	0.936
<b>TurboFold II</b>	0.766	0.903	0.770	0.925	0.770	0.925
<b>RNAalifold + SHAPE</b>	0.423	0.763	0.398	0.937	0.348	0.982
<b>RNAalifold</b>	0.395	0.831	0.395	0.937	0.348	0.982
<b>Rsample</b>	0.636	0.608	0.636	0.608	0.636	0.608
<b>MaxExpect</b>	0.773	0.850	0.773	0.850	0.773	0.850

SAM I riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.756	0.926	0.763	0.944	0.768	0.962
<b>TurboFold II</b>	0.769	0.959	0.769	0.962	0.769	0.959
<b>RNAalifold + SHAPE</b>	0.158	0.574	0.310	0.946	0.322	0.920
<b>RNAalifold</b>	0.158	0.574	0.106	0.485	0.115	0.435
<b>Rsample</b>	0.718	0.800	0.718	0.800	0.718	0.800
<b>MaxExpect</b>	0.718	0.800	0.718	0.800	0.718	0.800

M-box riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.874	0.916	0.875	0.913	0.875	0.913
<b>TurboFold II</b>	0.874	0.917	0.875	0.913	0.875	0.913
<b>RNAalifold + SHAPE</b>	0.733	0.876	0.550	0.875	0.616	0.971
<b>RNAalifold</b>	0.768	0.869	0.615	0.881	0.655	0.965
<b>Rsample</b>	0.875	0.913	0.875	0.913	0.875	0.913
<b>MaxExpect</b>	0.875	0.913	0.875	0.913	0.875	0.913

Lysine riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.840	0.985	0.840	0.984	0.841	0.988
<b>TurboFold II</b>	0.839	0.981	0.841	0.977	0.841	0.980
<b>RNAalifold + SHAPE</b>	0.417	0.774	0.356	0.886	0.244	0.849
<b>RNAalifold</b>	0.540	0.801	0.401	0.893	0.262	0.866
<b>Rsample</b>	0.810	0.894	0.810	0.894	0.810	0.894
<b>MaxExpect</b>	0.810	0.864	0.810	0.864	0.810	0.864

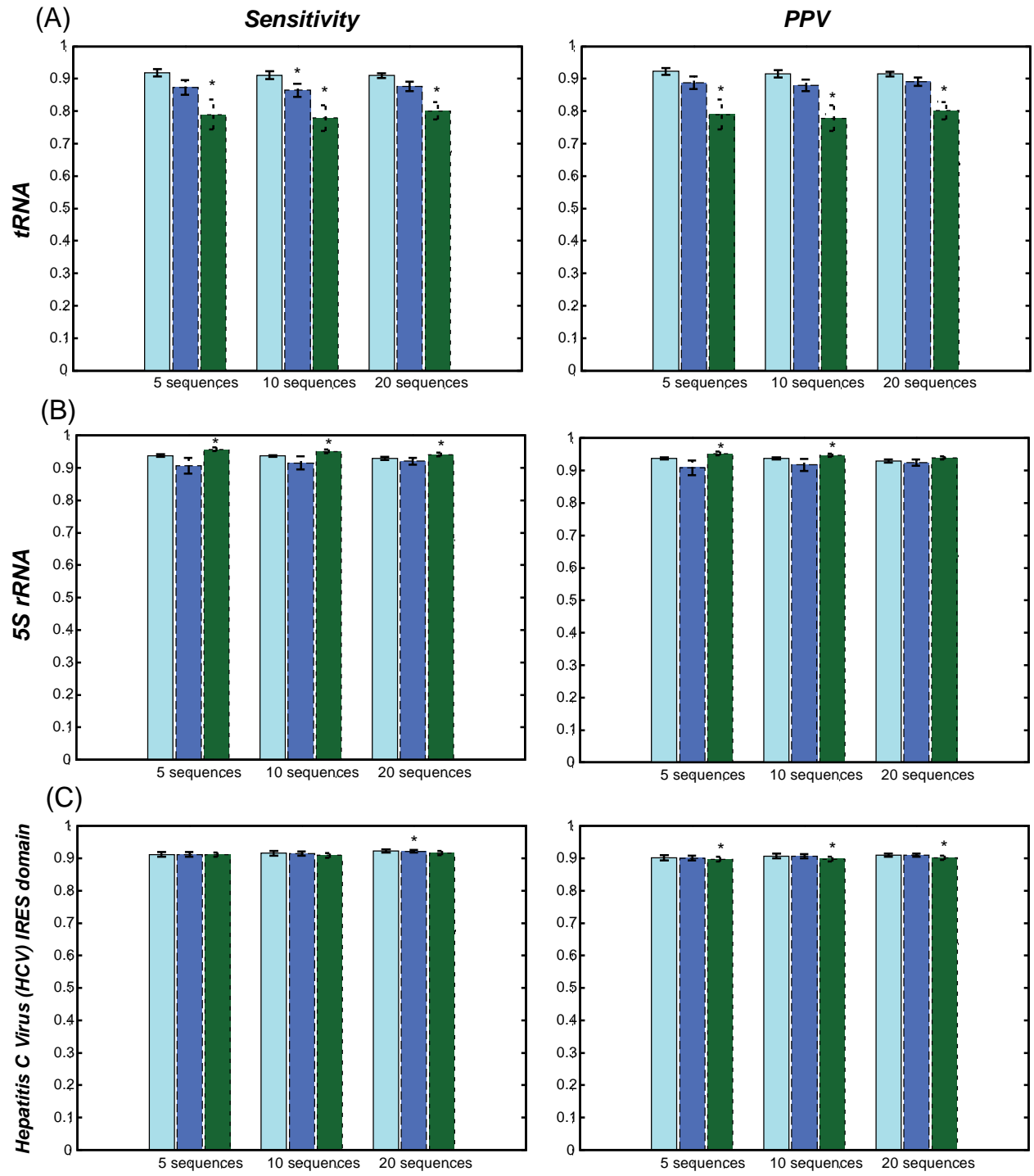
  




Cyclic-di-GMP riboswitch						
Prediction Method	$H = 5$ sequences		$H = 10$ sequences		$H = 20$ sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.802	0.860	0.761	0.824	0.788	0.821
<b>TurboFold II</b>	0.755	0.914	0.750	0.913	0.750	0.913
<b>RNAalifold + SHAPE</b>	0.498	0.792	0.492	0.937	0.395	0.964
<b>RNAalifold</b>	0.521	0.910	0.485	0.928	0.385	0.968
<b>Rsample</b>	0.929	0.928	0.929	0.928	0.929	0.928
<b>MaxExpect</b>	0.929	0.928	0.929	0.928	0.929	0.928

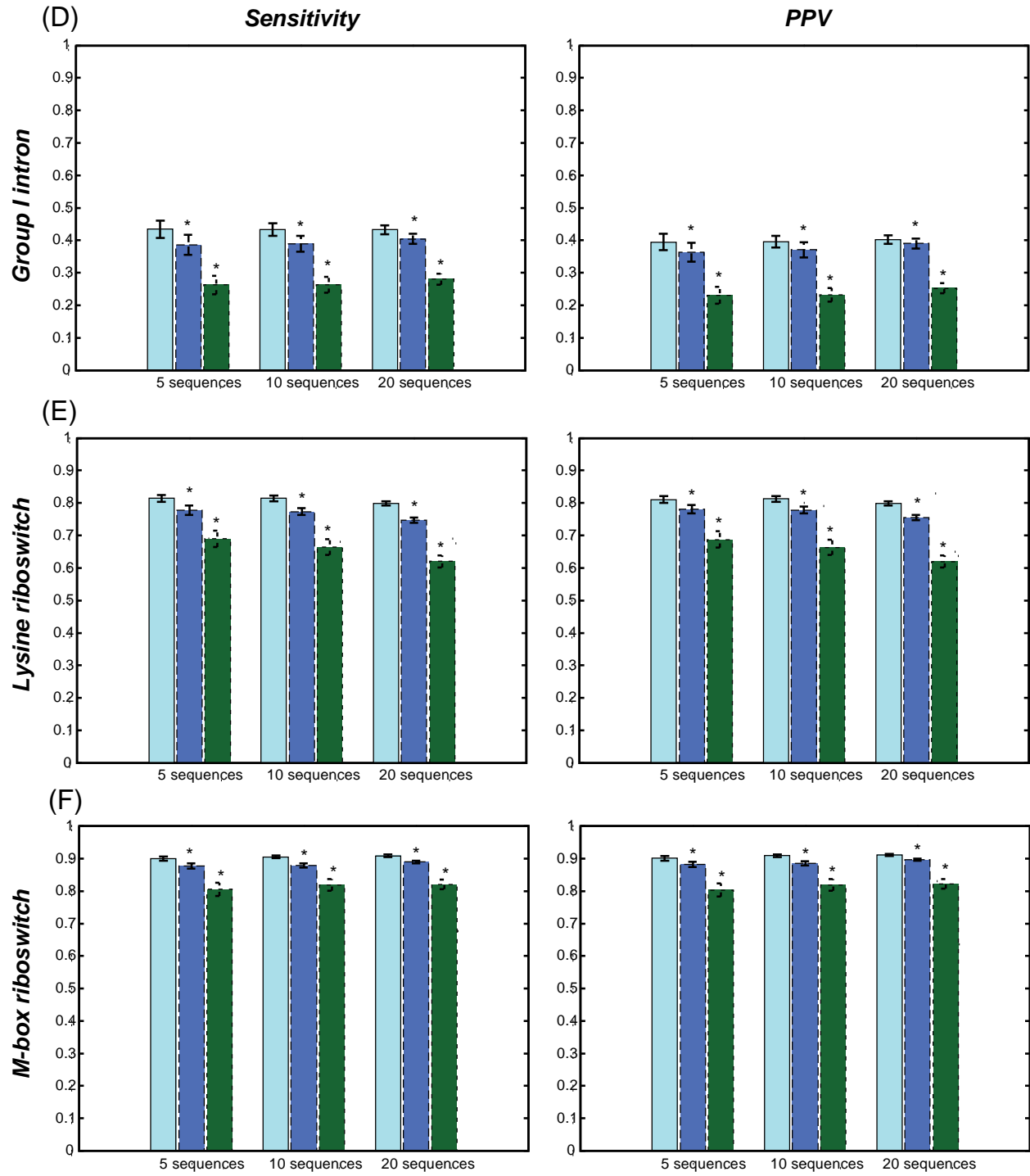


<b>23S rRNA</b>						
Prediction Method	<i>H</i> = 5 sequences		<i>H</i> = 10 sequences		<i>H</i> = 20 sequences	
	sensitivity	PPV	sensitivity	PPV	sensitivity	PPV
<b>TurboFold II + SHAPE</b>	0.879	0.866	0.888	0.869	0.863	0.849
<b>TurboFold II</b>	0.834	0.831	0.867	0.862	0.878	0.869
<b>RNAalifold + SHAPE</b>	0.730	0.789	0.732	0.870	0.737	0.901
<b>RNAalifold</b>	0.797	0.823	0.786	0.887	0.758	0.905
<b>Rsample</b>	0.713	0.692	0.713	0.692	0.713	0.692
<b>MaxExpect</b>	0.605	0.576	0.605	0.576	0.605	0.576

*TurboFoldII +SHAPE* ■  
*TurboFoldII* ■  
*ClustalW* ■



*TurboFoldII +SHAPE*   
*TurboFoldII*   
*ClustalW* 



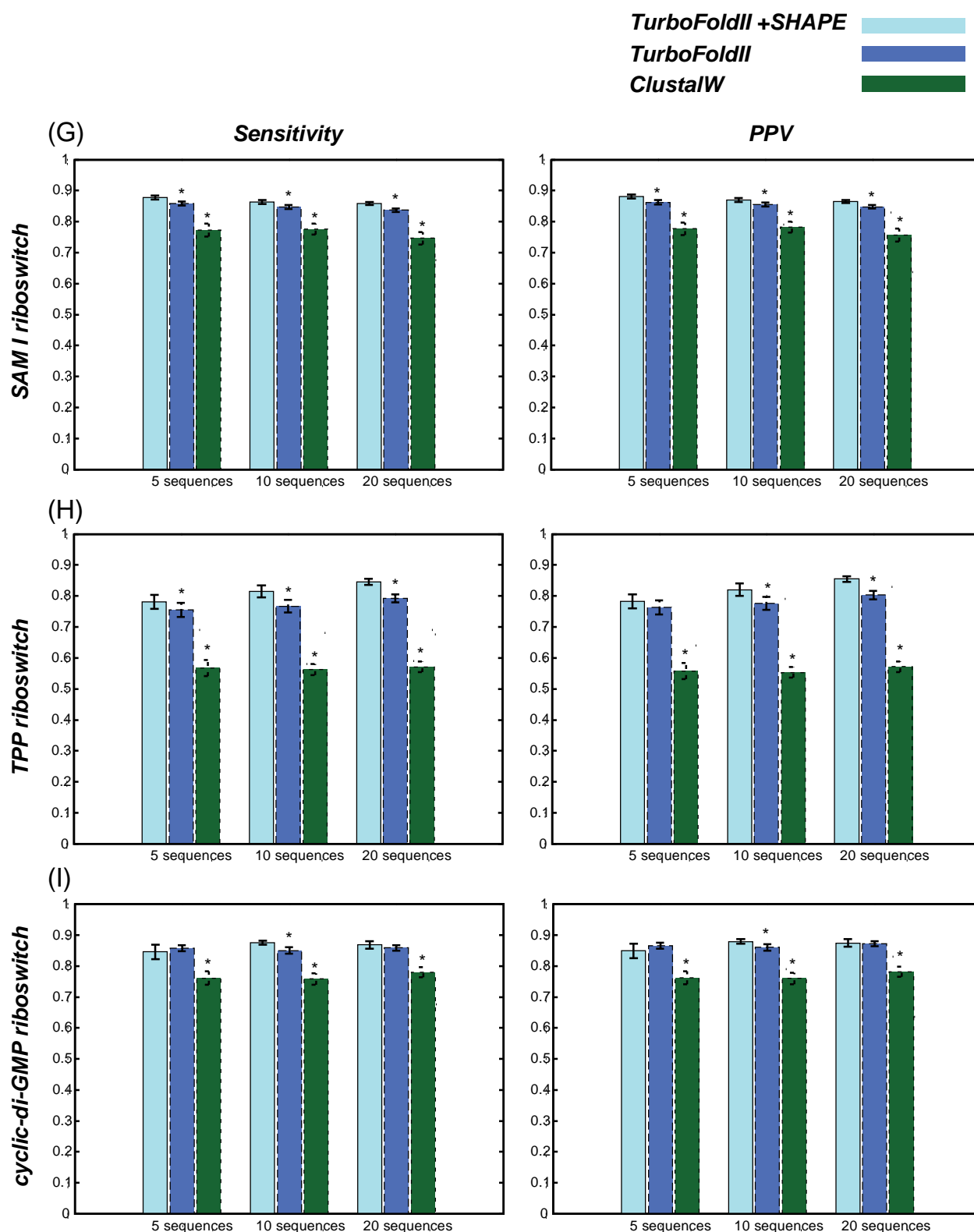
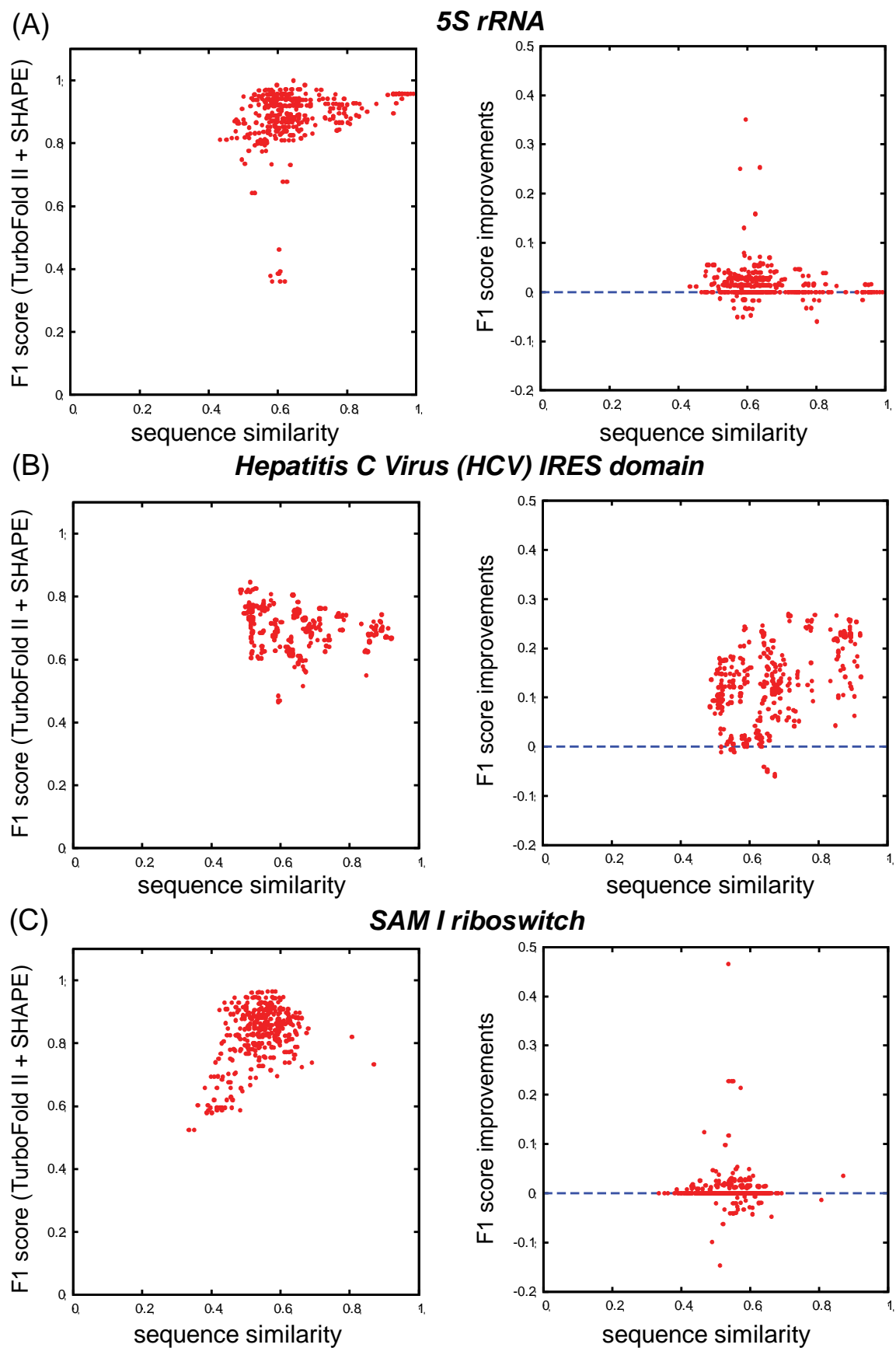


Figure S13. **Sensitivity and PPV of multiple sequence alignment of sequences that without SHAPE mapping data.** Sensitivity and PPV of multiple sequence alignment obtained by running the methods with 5, 10, or 20 input sequences on (A) tRNA, (B) 5S rRNA, (C) hepatitis

C virus IRES domain, (D) group I intron, (E) lysine riboswitch, (F) M-box riboswitch, (G) SAM I riboswitch, (H) TPP riboswitch, and (I) cyclic-di-GMP riboswitch RNA test datasets. The star (\*) above the bar for a method indicates that the difference in sensitivity (left) and PPV (right) between the method and TurboFold II+SHAPE is statistically significant, as determined by paired t-tests.



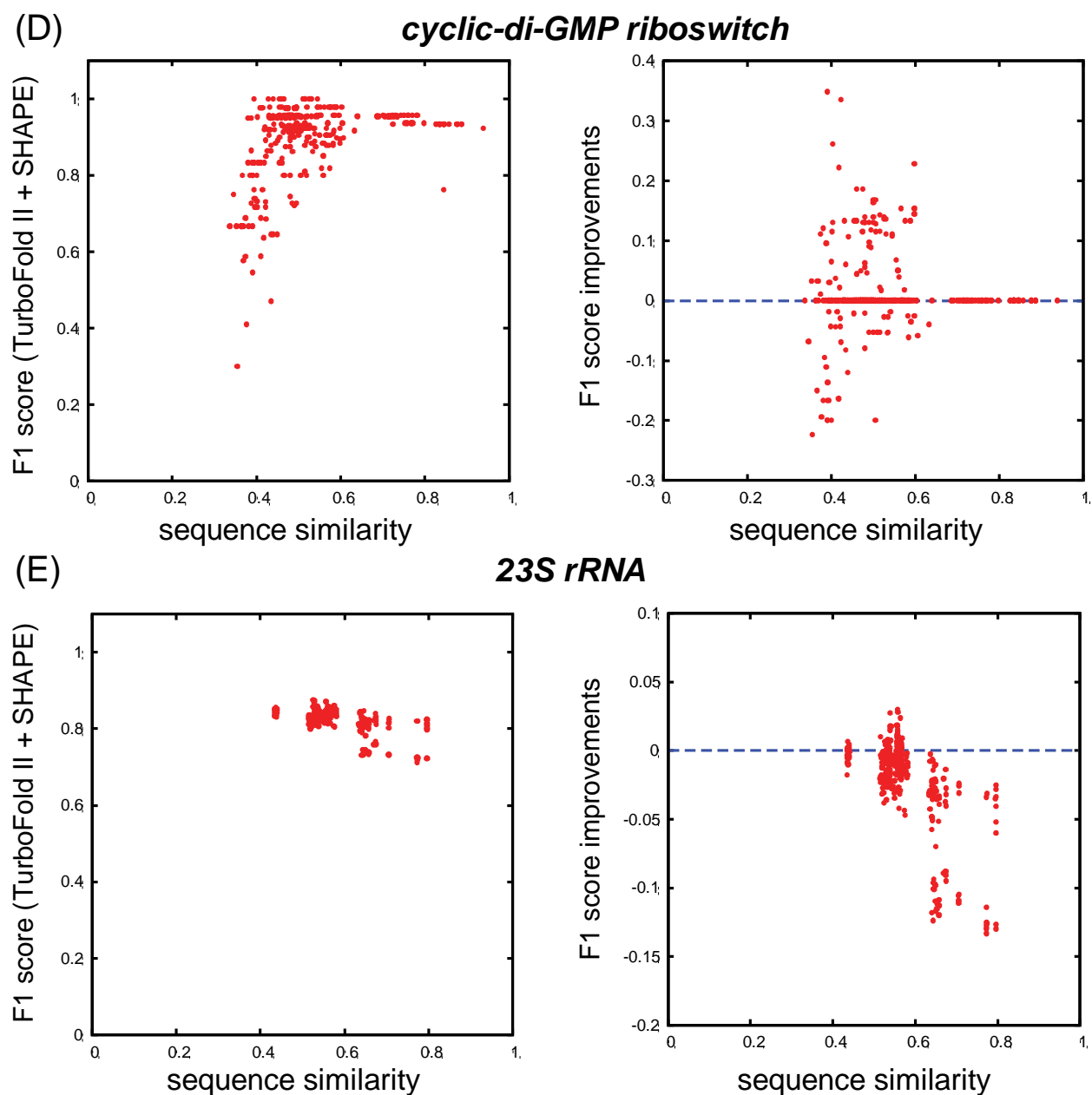


Figure S14. Scatter plots of F1 score of structure prediction obtained with TurboFold II + SHAPE prediction as a function on sequence similarity (left). Difference between F1 scores of structure prediction obtained with TurboFold II + SHAPE and TurboFold II as a function on sequence similarity (right). Sequences from (A) 5S rRNA, (B) Hepatitis C Virus (HCV) IRES domain, (C) SAM I riboswitch, (D) cyclic-di-GMP riboswitch and (E) 23S rRNA datasets.

## Section 2. Parameter optimization methods

The extrinsic information for nucleotides  $i$  and  $j$  in sequence  $m$  is represented as

$$P^{(n \rightarrow m)}(i, j) = \sum \begin{cases} \sum_{\substack{1 \leq k < l \leq N_n \\ k \in C_i^{m,n} \\ l \in C_j^{m,n}}} Prob_{bp} \times P^{(m,n)}(i \sim k) \times P^{(m,n)}(j \sim l) \times (H - 1) \times \lambda & \text{(if sequence } n \text{ is with SHAPE)} \\ \sum_{\substack{1 \leq k < l \leq N_n \\ k \in C_i^{m,n} \\ l \in C_j^{m,n}}} Prob_{bp} \times P^{(m,n)}(i \sim k) \times P^{(m,n)}(j \sim l) \times (1 - \psi_{m,n}) & \text{(otherwise)} \end{cases} \quad (1)$$

To train the parameter  $\lambda$ , 20 groups of input sequences formed by 10 homologous sequences (including the sequence with SHAPE data) were randomly chosen from the small subunit ribosomal RNA in the database RNAStralign. The range for parameter  $\lambda$  is from 0 to 2.0 (0, 0.02, 0.1, 0.2, 0.4, 1.0, 1.6, 2.0). The resulting optimal parameters ( $\lambda = 1.0$ ) was then used as the default for the method.

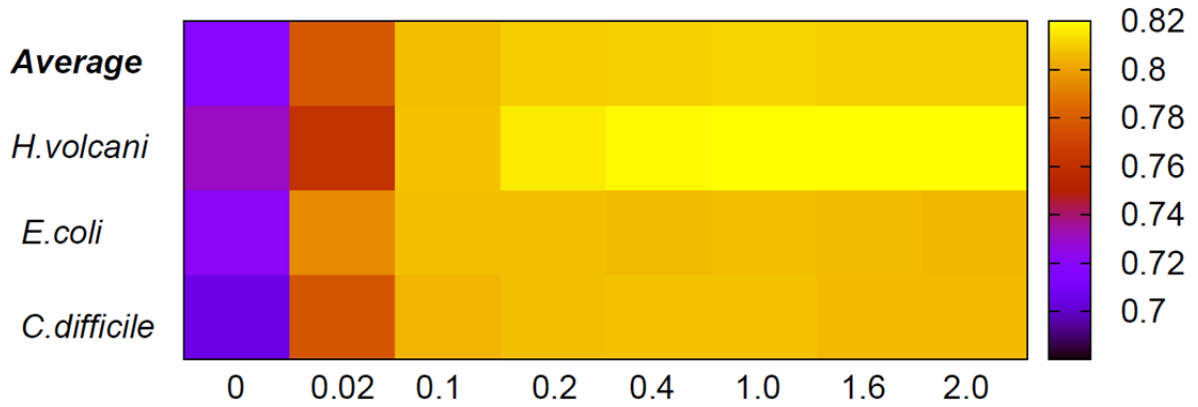


Figure S15. **Grid search plots for parameters  $\lambda$  in extrinsic information calculation.** The scale of heat map shows that the grids with higher values are represented in lighter color.



### Section 3. Software efficiency test

Table S6. **Run time of the program over the selected families and different number of input sequences.** The calculations were run on one core on a machine with an Intel® Core™ i7-4790 CPU @ 3.60GHz. Except the tests of 23S rRNA were run by parallel using 6 cores, other tests were on single core.

Family	<i>H=5 sequences</i>		<i>H=10 sequences</i>		<i>H=20 sequences</i>	
	TurboFold II+SHAPE	TurboFold II	TurboFold II+SHAPE	TurboFold II	TurboFold II+SHAPE	TurboFold II
<i>tRNA</i>	13.00s	8.65s	29.17s	24.12s	1m28s	1m24s
<i>Group I intron</i>	3m50s	3m4s	12m51s	12m9s	35m4s	32m41s
<i>23S rRNA</i>	3h47m	3h46m	7h46m	7h45m	21h4m	20h6m

### Section 4. Sequences used in parameter optimization and benchmarking

16S ribosomal RNA sequences were separated into groups of 10 input sequences for each family for parameter optimization. Sequences were also separated in 5, 10 and 20 input sequences for benchmarking: 5S ribosome RNA, 23S ribosome RNA, tRNA, hepatitis C virus (HCV) IRES domain, group I intron, lysine riboswitch, M-box riboswitch, SAM I riboswitch, TPP riboswitch, and cyclic-di-GMP riboswitch. Specific sequences lists are provided in “TurboFoldII\_SHAPE\_SelectedSequenceLists.zip”.