Designing molecular RNA switches with Restricted Boltzmann machines

SUPPLEMENTAL MATERIAL

Jorge Fernandez-de-Cossio-Diaz, Pierre Hardouin, Francois-Xavier Lyonnet du Moutier, Andrea Di Gioacchino, Bertrand Marchand, Yann Ponty, Bruno Sargueil, Rémi Monasson, Simona Cocco

Appendix A: Further details on RBM implementation

1. RBM training

The RBM is trained using the ADAM [1] optimization algoirthm, with minibatches containing 128 sequences and for a total of 10000 gradient update steps. At each step, the gradient is computed by Persistent Contrastive Divergence (PCD) [2], taking 100 Monte Carlo steps to update the Markov chains in each iteration. The number of Monte Carlo chains equals 128. During training the pseudolikelihood is monitored to assess convergence. The gradient is centered as described in [3]. In addition, hidden units are re-scaled to a variance of 1 at each iteration by a simple scaling transform. The implementation is identical to the one used in [4].

2. RBM sampling

The RBM is sampled by the Gibbs algorithm [5], by alternating between sampling the visible configuration conditional on the current hidden state, and sampling a new hidden state conditional on the current visible configuration. Because of the bipartite architecture of the RBM, these conditional distributions factorize and are easy to sample.

To ensure samples are approximately in equilibrium, we monitor the effective eneergy of the sample, E_{eff}. In our experiments we found that 10000 Gibbs sampling steps were more than sufficient to reach a stable plateau. For more implementation details we refer to [4].

3. Cross-validation analysis

The RBM employed in the main-text has 100 dReLU hidden units and was trained with a regularization penalty of $\lambda_{\text{reg}} = 0.01$. After extensive exploration of other numbers of hidden units and regularization penalties, we chose these settings as a good balance between model quality (as measured by the log-likelihood of withheld data) and simplicity. Figure S1 shows the results of the experiments performed at different regularization penalties (panel A) and numbers of hidden units (panel B). Beyond 100 hidden units there are dimishing returns in the validation log-likelihood.

We have conducted cross-validation tests of our model to control for overfitting. We trained new RBMs on reduced datasets, consisting of 80% randomly selected sequences of the MSA for training, and 20% withheld for validation. We then computed the scores $-E_{\text{eff}}$ of sequences used in training and validation, and compared their histograms in S1C. Both histograms are in close agreement, indicating that our hyper-parametric choices place the model far from an overfitted regime. To further justify our architectural choices in the RBM, we experimented various choices for the number of hidden units of the RBM and the regularization strength λ_{reg}. Regarding the regularization strength, we find that for λ_{reg} larger than 0.05, we suffer a loss of log-likelihood of the validation data. On the other hand, for λ_{reg} smaller than 0.01 there is no significant gain in likelihood. The results of these experiments are shown in Figure S1A. We have then set $\lambda_{\text{reg}} = 0.01$, since a non-zero regularization makes the model more robust in case of very conserved sites, and makes the model easier to sample and train [5, 6]. The hyper-parameter scan also reveals diminishing returns for more than 100 hidden units in terms of the log-likelihood of validation data, as shown in Figure S1B. Therefore, we have chosen to use 100 hidden units in this work.

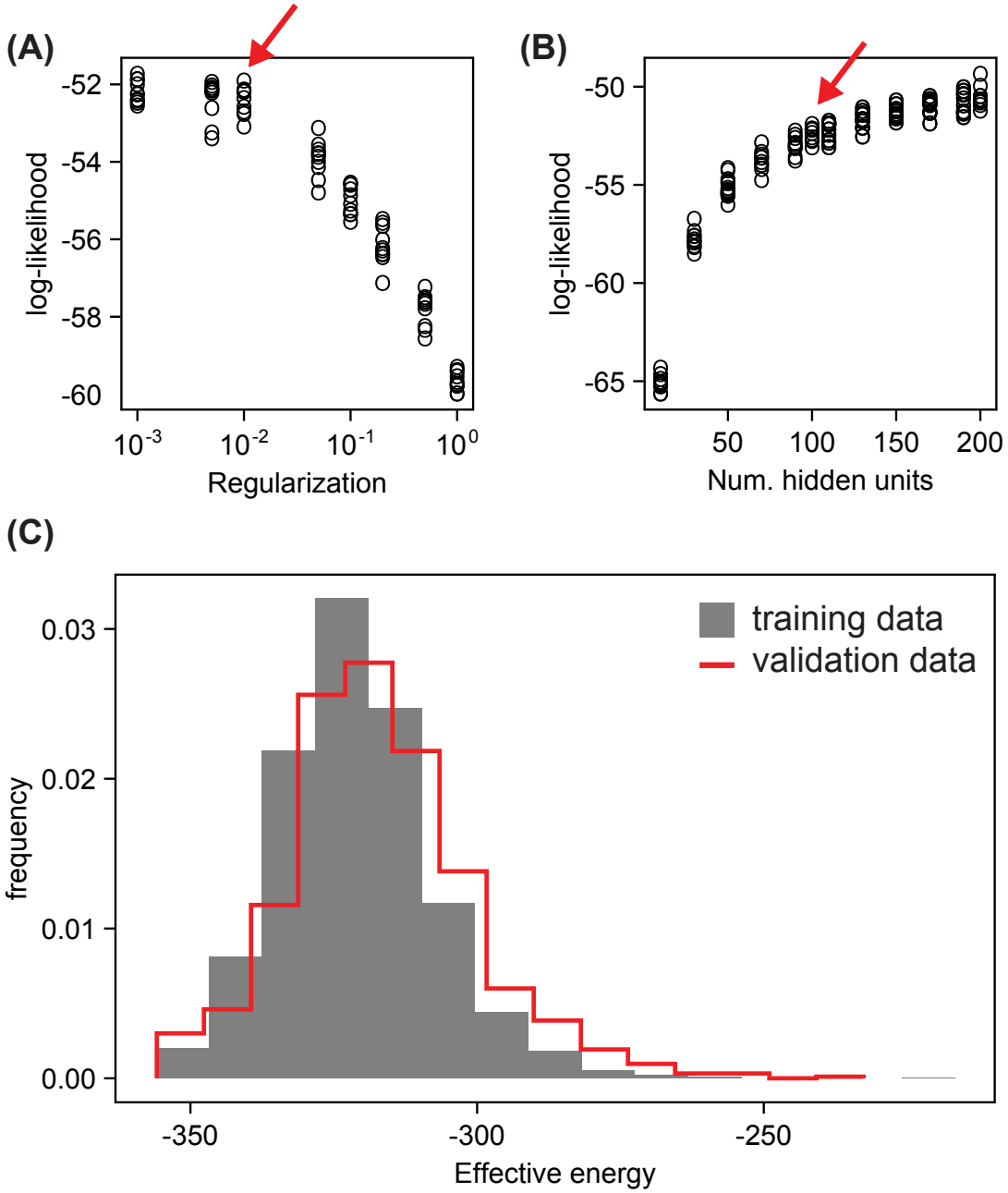

FIG. S1. Validation of RBM architectural choices. **(A)** RBMs were trained with different regularization weights, and **(B)** different numbers of hidden units (from 20 to 200), on different random subsamples containing 80% of the training data. The plots show the log-likelihood of the remaining 20% of validation data. Redd arrows in D,E mark our final choices for the regularization ($\lambda_{\text{reg}} = 0.01$) and number of hidden units ($M = 100$). **(C)** An RBM with 100 hidden units and $\lambda_{\text{reg}} = 0.01$, trained on a randomly selected subset containing 80% of training sequences was used to compute the effective energies of sequences in its training set and the remaining 20% of sequences in the validation set. Histograms of the effective energies, E_{eff} from (5), of the two groups (training data in gray, validation data in red) of sequences shows excellent agreement, suggesting no overfitting has occurred.

Appendix B: Reactivity histograms

Figure S2 shows the histograms of reactivities of paired and unpaired sites, under some modifications of the criteria used to produce Fig. 6A in the main-text. Overall, the histograms produced are similar, and thus robust, to these variations. We have also attempted to assess the impact of experimental noise (*e.g.*, from the finite number of sequencing reads) on the observed histograms in Fig. S3. Although we cannot directly access the idealized reactivities at infinite read depth, we can go in the opposite direction and resample the reactivities according to the distributions distributions $P_{ni}(r|\tilde{r}_{ni})$ to evaluate the effect adding *more* noise. If the histograms do not change significantly, this suggests that the noise, though present, might not play a significant role in affecting our conclusions. Figure S3 shows the resampled reactivity histograms of paired and unpaired sites (using the same site classification as in Fig. 6A of the main text). Although there are some small differences, overall the resampled histogram mostly overlaps with the original histograms from the reactivity dataset.

Figure S4 shows the reactivity histograms of the sites belonging to the P1 helix in two conditions: without SAM (left) and with SAM (right). In the same fashion as the pseudoknot (cf. Fig. 6B in the main text), the P1 reactivities shift from a distribution compatible with unpaired residues in absence of SAM, to a distribution closer to the paired sites reactivities, in the presence of SAM. This result supports the notion that P1 is stabilized when SAM is bound. We remark that, in comparison to the pseudoknot (Fig. 6B in the main text), P1 sites in presence of SAM exhibit slightly larger reactivities, indicating that some aptamers might not fully base-pair along this helix in response to SAM. Like Fig. 6B in the main text, this result also supports the validity of the histograms in Fig. 6A as accurate approximations of the statistical reactivities of paired and unpaired sites.

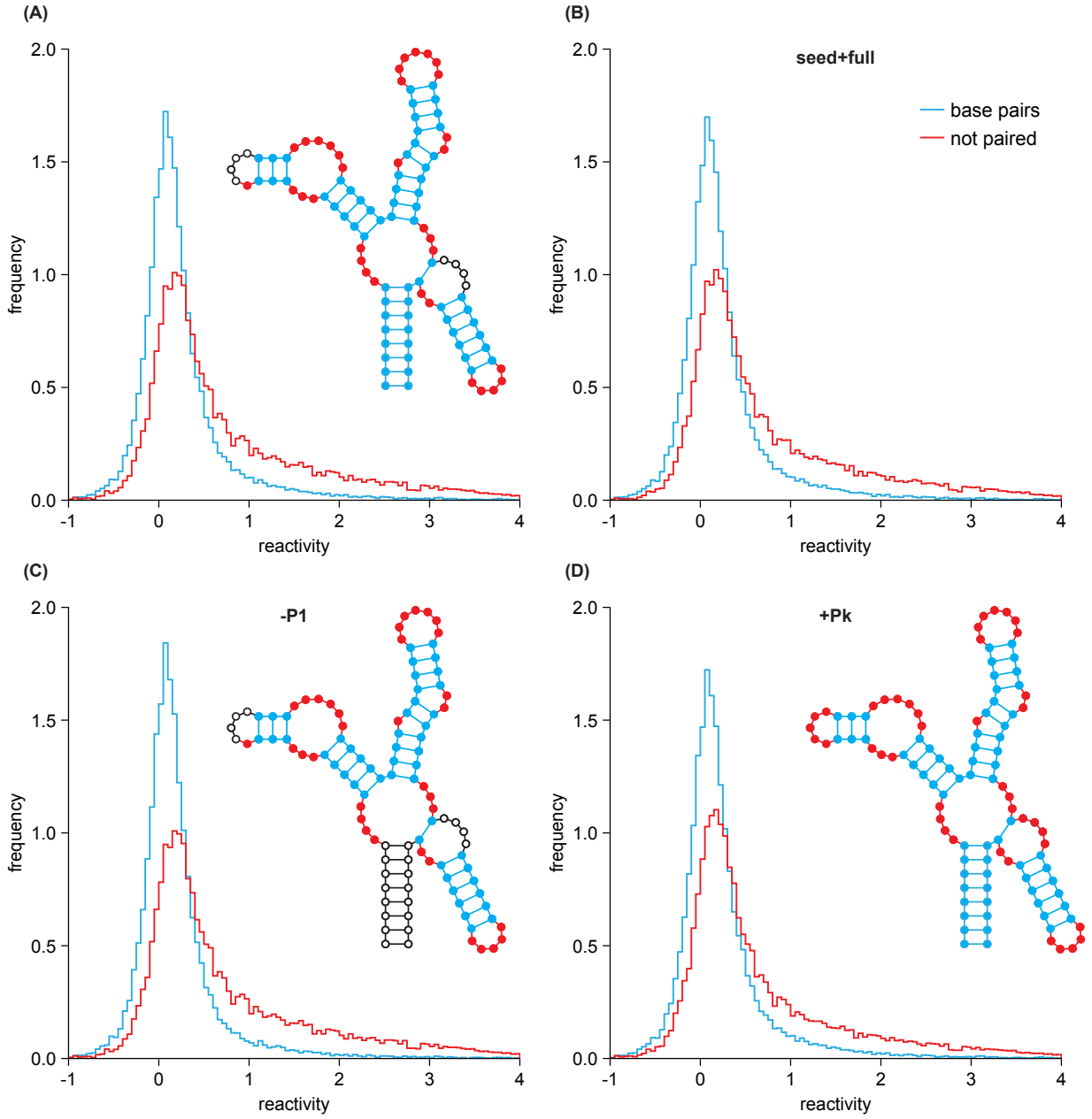

FIG. S2. Histograms of paired and unpaired site reactivities, according to different criteria. **(A)** Same as Figure 6A of the main text. **(B)** Like (A), but including all natural sequences (seed MSA + full MSA). **(C)** Like (A), but excluding the P1 helix from the set of base-paired sites. **(D)** Like (A), but including the sites involved in the pseudoknot in the set of unpaired sites. In the four panels, the inset secondary structure indicates the sites included as base-paired (blue), as unpaired (red), and sites excluded from both histograms (unfilled circles).

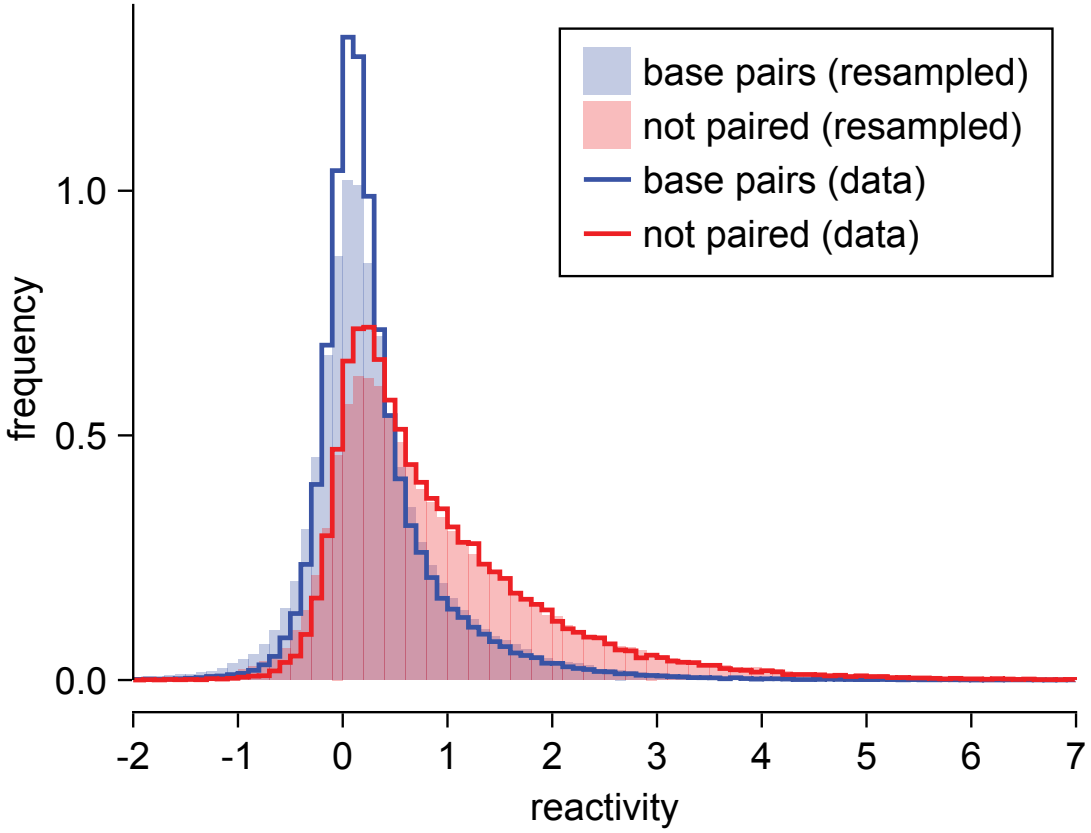

FIG. S3. Histograms of paired and unpaired site reactivities, after resampling reactivities from the site distributions $P_{ni}(r|\tilde{r}_{ni})$ defined in the main-text. The original histograms are shown in continuous line, while the resampled histograms are filled.

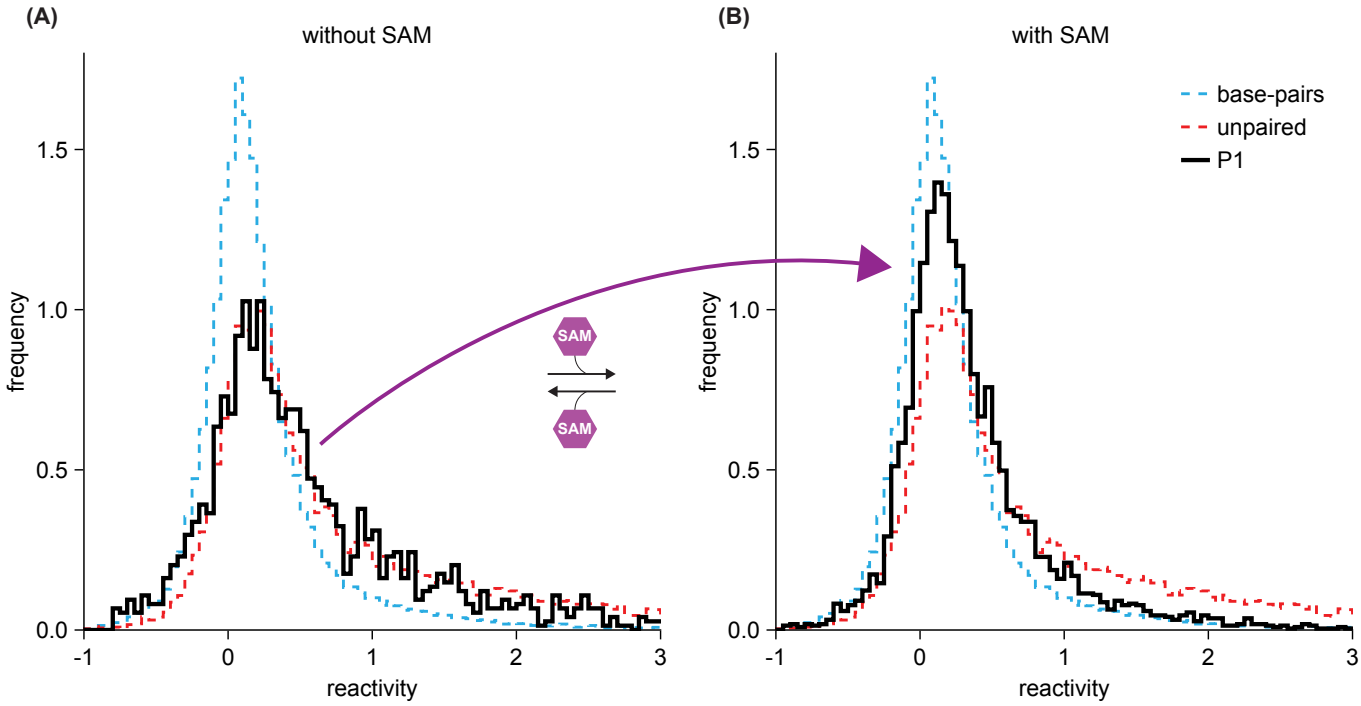

FIG. S4. Stabilization of P1 formation is reflected in SHAPE reactivity histograms. **(A)** Histogram of reactivities of base-paired (blue) and unpaired sites (red) in probed natural sequences belonging to the manually curated seed alignment. Histogram of P1 sites, in absence of SAM, is shown in black, and agrees with the distribution of unpaired sites in absence of SAM. **(B)** In presence of SAM, the reactivities of P1 sites move towards a distribution compatible with base-pairing.

Appendix C: Principal components analysis

We performed a principal component analysis (PCA) on the sequences of the full RF00162 multiple-sequence alignment (MSA), as explained in the main text (Methods). Figure S5 shows the top two components, in sequence logo representation. The first component notably reflects the deletion of the P4 helix in a cluster of natural sequences, mostly Actinomycetota. This is appreciated from the prominence of gaps in sites from 81 to 99. See also Fig. 4B in the main text. Thus the top component separates sequences in which P4 is present (and thus have a positive projection onto this component) from sequences in which P4 is delete (having a negative projection onto this component). Various reports in the literature have discussed the role of P4 in the function of the riboswitch. In particular, [7] found sequneces without P4 able to bind SAM, and were able to correlate experimentally the length of P4 with SAM affinity.

Figure S6 shows the projection of the sequences probed in the first experiment along the two top principal components (PC), annotated by our protection score as switcher (filled circles) or non-switcher (empty circles). As can be seen in the figure, probed sequences span the space defined by these two PCs, suggesting adequate coverage of the diversity represented in the MSA.

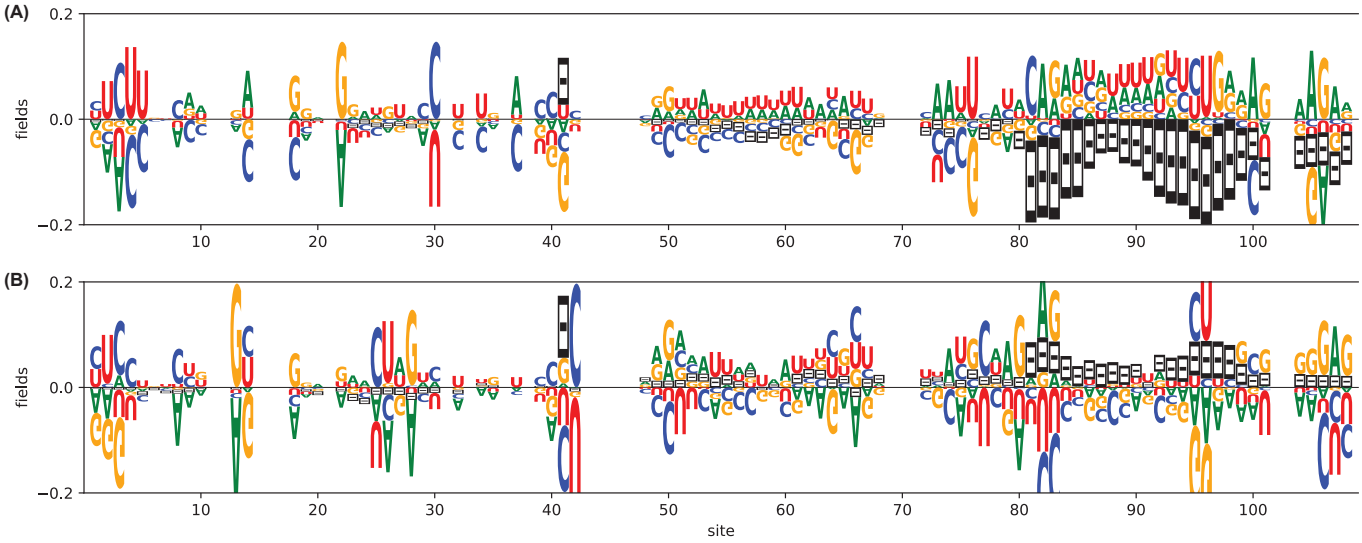

FIG. S5. **(A)** Sequence logo of the eigenvector corresponding to the largest eigenvalue of the correlation matrix of the RF00162 MSA. **(B)** Sequence logo of the eigenvector corresponding to the second largest eigenvalue of the correlation matrix of the RF00162 MSA.

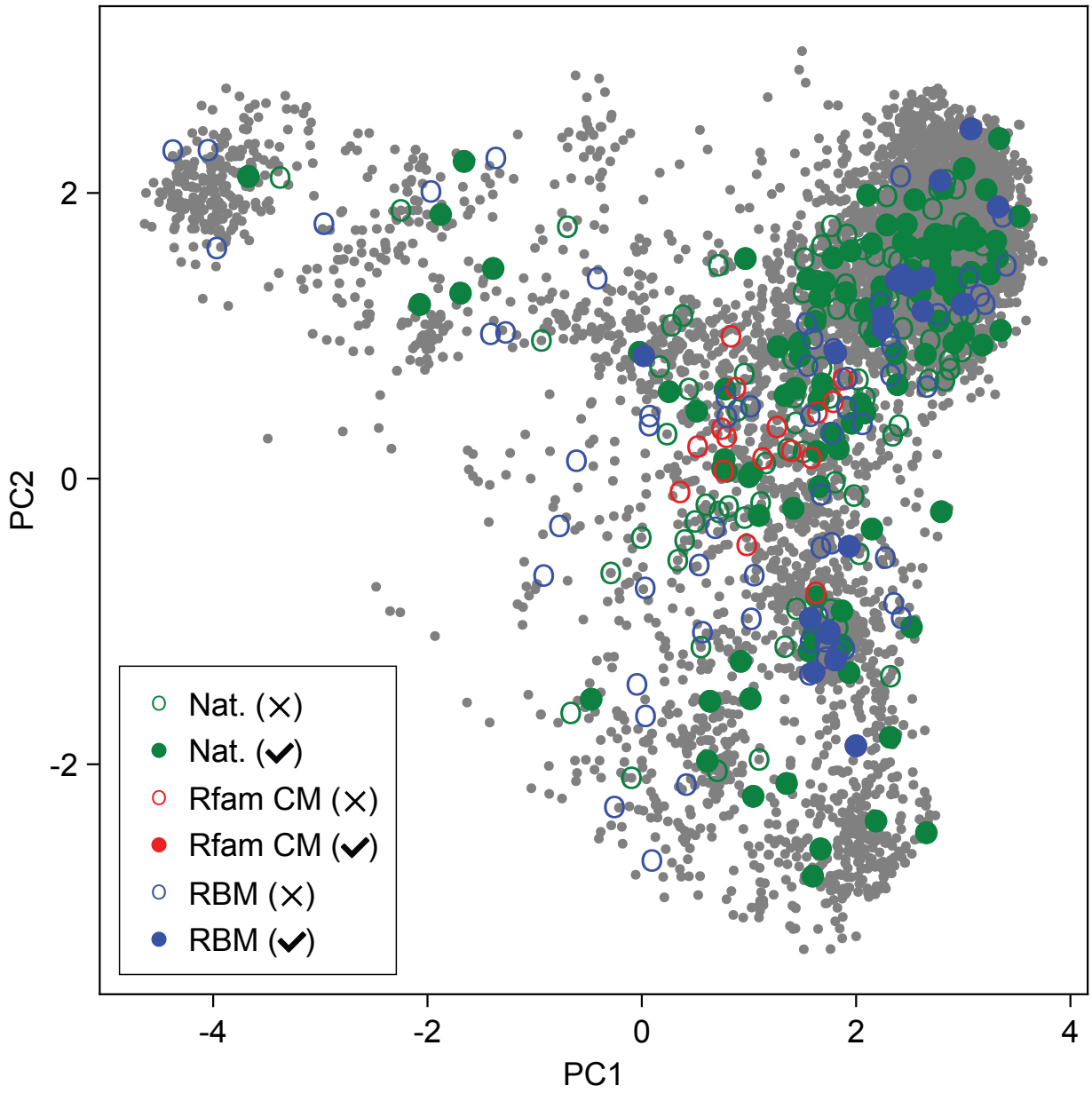

FIG. S6. Projection of probed sequences onto principal components of the RF00162 MSA. Sequences are colored by their origin: Natural (green), Rfam CM (red), and RBM (blue). Sequences that respond to SAM are shown as filled disks, whereas unresponsive sequences are shown as empty circles.

Appendix D: Supplementary average reactivity profiles

Rfam alignments are constituted of a manually curated alignment of "seed" sequences (which consists of 457 entries for the RF00162 family), and in addition a set of "hits", fetched from genome databases by Infernal as having high bit-score with the Rfam CM. It is plausible that some of the hits could be false positives and thus not true aptamers. In Fig. S7, we show that the average reactivity differences in response to SAM of the full MSA is in excellent agreement to the average reactivities differences of the seed MSA. This suggests that the set of natural sequences in our experiments are able to bind SAM, and respond by performing the expected structural switch.

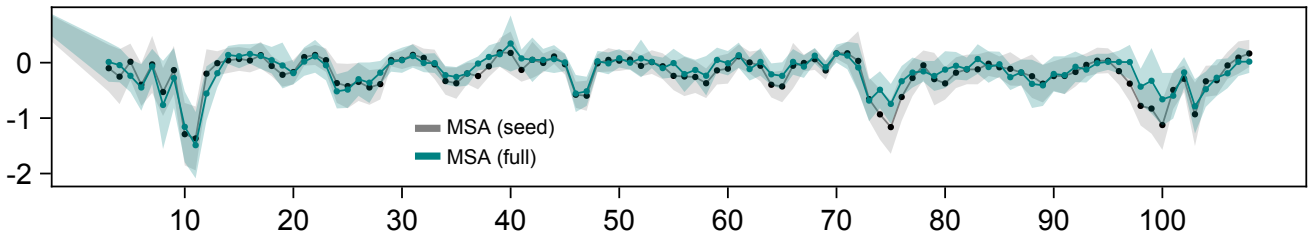

FIG. S7. Average differential reactivities in response to SAM. We compute the reactivity difference of the Mg+SAM condition minus the Mg condition for probed sequences from the full MSA (teal). The plot shows the profile averaged over sequences, with bands showing the ± one standard deviation. In gray, we show the same plot but for seed sequences.

We also performed experiments in absence and presence of magnesium (Mg), which is known to stabilize the secondary structure of RNA. In Figure S8, we show the average reactivity response to magnesium (Mg) of seed alignment natural sequences, RBM of high and low scores (> 300 and < 300, respectively), and of the Rfam CM generated sequences. The reactivity responses to Mg of RBM designed sequences of high score are in excellent agreement with those of the natural sequences (Pearson correlation $= 0.89$). In contrast, RBM generated sequences of low RBM scores, or Rfam generated sequences, exhibit larger deviations (Pearson correlations of 0.68 in both cases).

As mentioned in the text, RBM generated sequences of low score also exhibit larger deviations in their reactivity responses to SAM in comparison to the natural sequences. These comparisons are shown in Fig. S9.

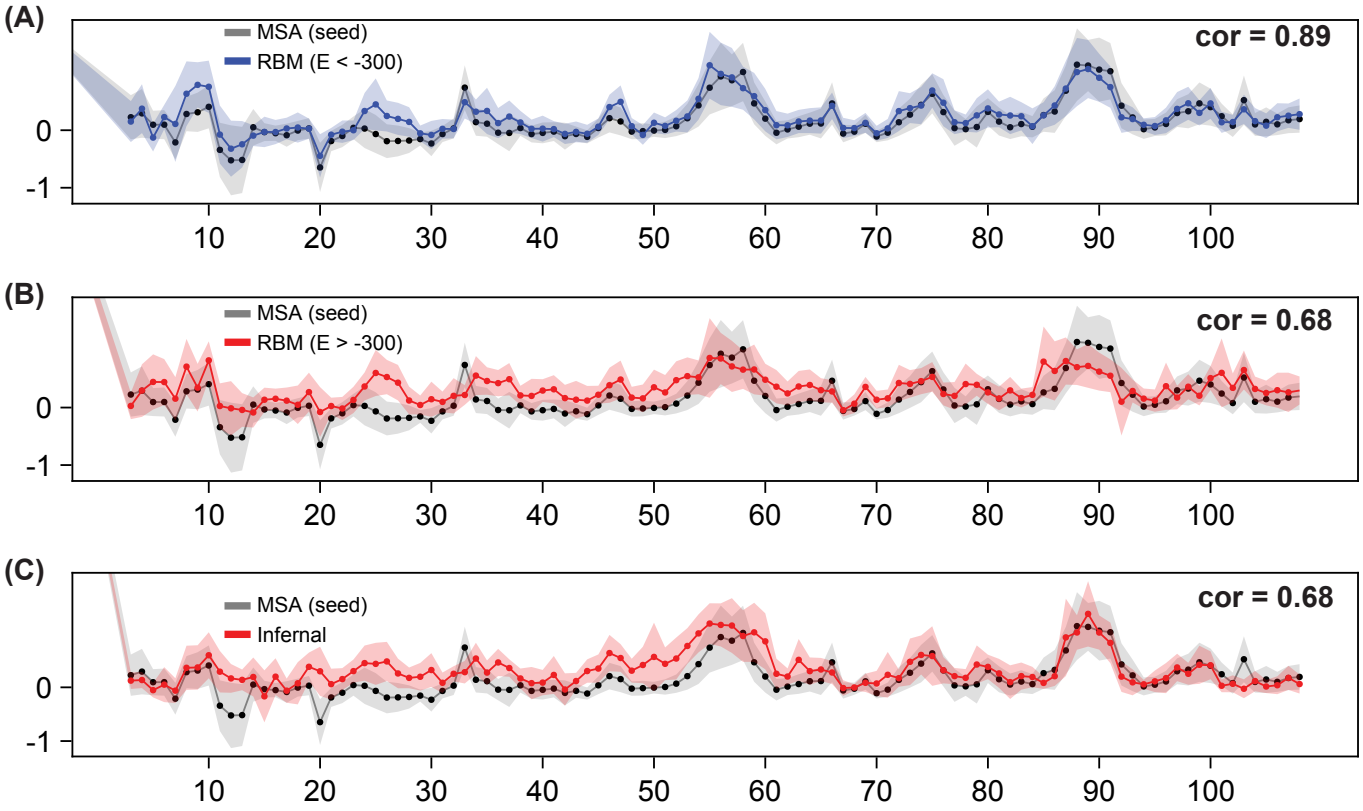

FIG. S8. Average differential reactivities in response to magnesium (Mg). We compute the reactivity difference of the Mg condition minus the 30C (no ligand) condition for different groups of probed sequences. **(A)** Shows the natural seed sequences (in gray) and the RBM sequences with low $E_{\text{eff}} < -300$. The plot shows the profile averaged over sequences, with bands showing the $\pm$ one standard deviation. In gray, we show the same plot but for seed sequences, as in Figure **??** of the main text. Then panels **(B)** and **(C)** are are analogous but for RBM sequences with high $E_{\text{eff}} > -300$, and for Rfam CM sequences, respectively. The Pearson correlations to the seed natural sequence reactivity profiles are indicated.

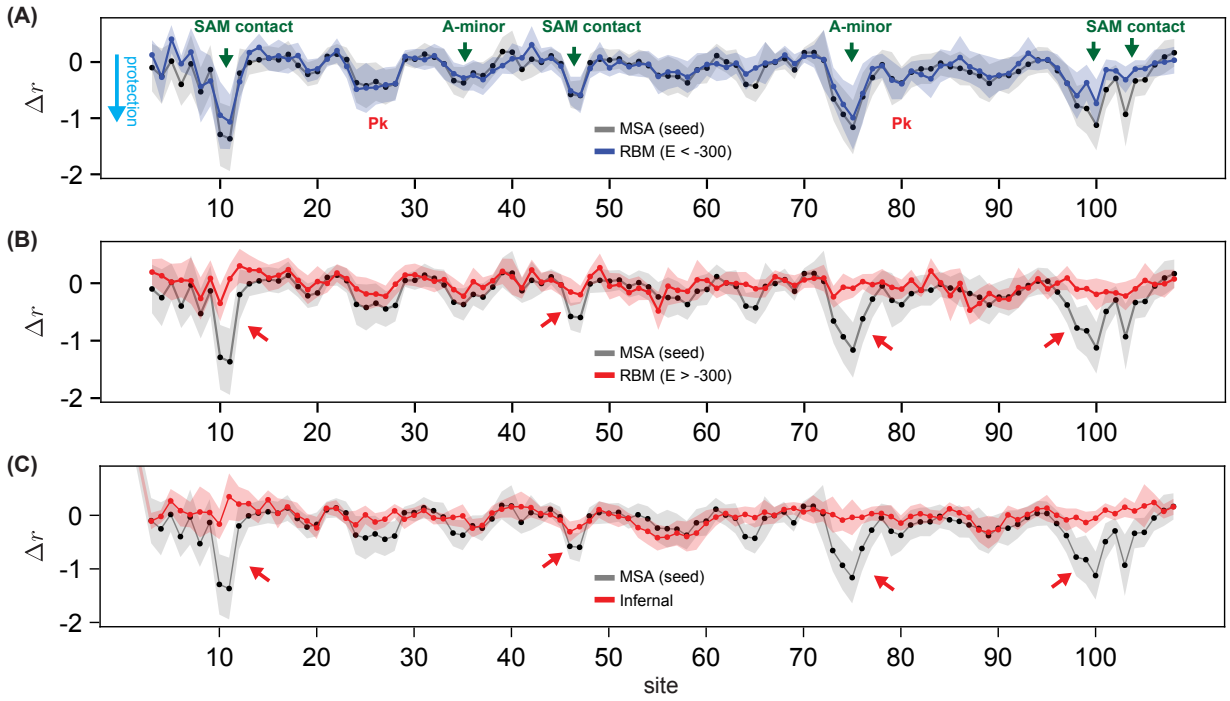

FIG. S9. Average differential reactivities in response to SAM: **(A)** for natural sequences (in gray) and low-E_{eff} RBM generated sequences (in blue; $E_{\text{eff}} < -300$), **(B)** for natural sequences (in gray) and low-E_{eff} RBM generated sequences (in red; $E_{\text{eff}} > -300$), and **(C)** for Infernal generated sequences, using the Rfam CM. For each group of sequences and each site, we computed the average reactivity difference of the condition with SAM+Mg minus the condition with Mg only. The plots show the average profiles of sequences in the group, with the bands indicating ± half one standard deviation. The green arrows indicate locations expected to exhibit reactivity differences in response to SAM binding, while the red arrows highlight lack of protection in some groups of sequences in those locations.

Appendix E: Further analysis of the Rfam, Denoised and Unknotted CM variants

In order to be able to capture more distant sequences in the alignments, the Rfam CM models are regularized, resulting in a less constrained model. We therefore considered two CM variants: Denoised and Unknotted CMs (see Methods in the main text for precise definitions). The regularization of Rfam CM manifests in the fact that it does not fit precisely the conservation profile of natural sequences, as shown in Fig. S10 (second row). On the other hand, we have checked that Denoised and Unknotted CMs fit accurately the MSA conservation profiles, Fig. S10 (third and fourth rows).

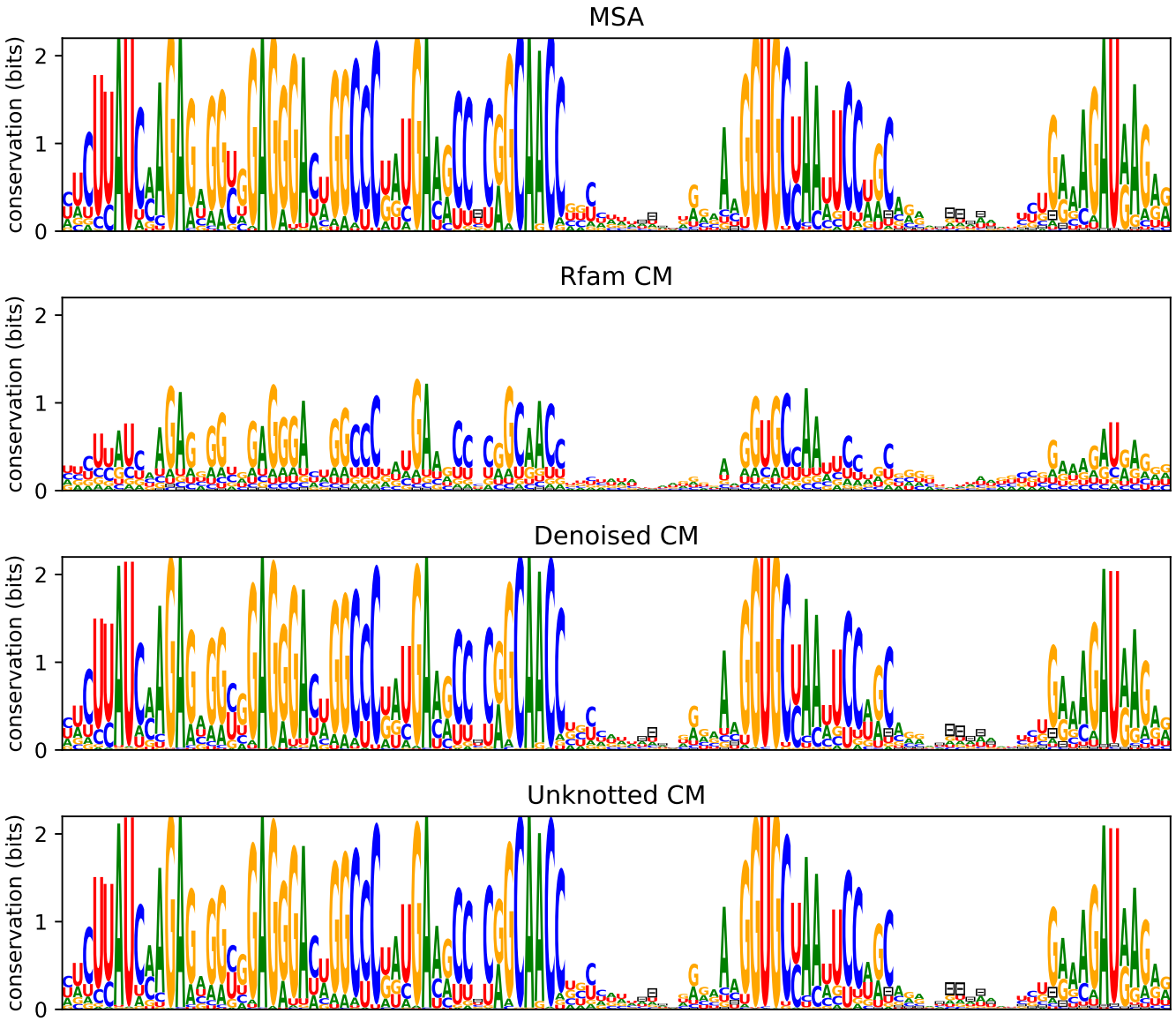

FIG. S10. Sequence logos of sequences sampled from the various CM variants defined in the text. The first row is the sequence logo of the natural MSA. The following rows, correspond to the Rfam CM, the Denoised CM, and the Untangled CM.

As we saw in Figure 4A in the main-text, the Rfam CM is unlikely to generate sequences of high RBM scores. We corroborate here that also the Denoised and Unknotted CMs are unlikely to sample sequences of high RBM score. We repeated the analyses of Figure 4 from the main-text, replacing the Rfam CM with the Denoised and Unknotted CMs. Results are shown in Supplementary Figure S11).

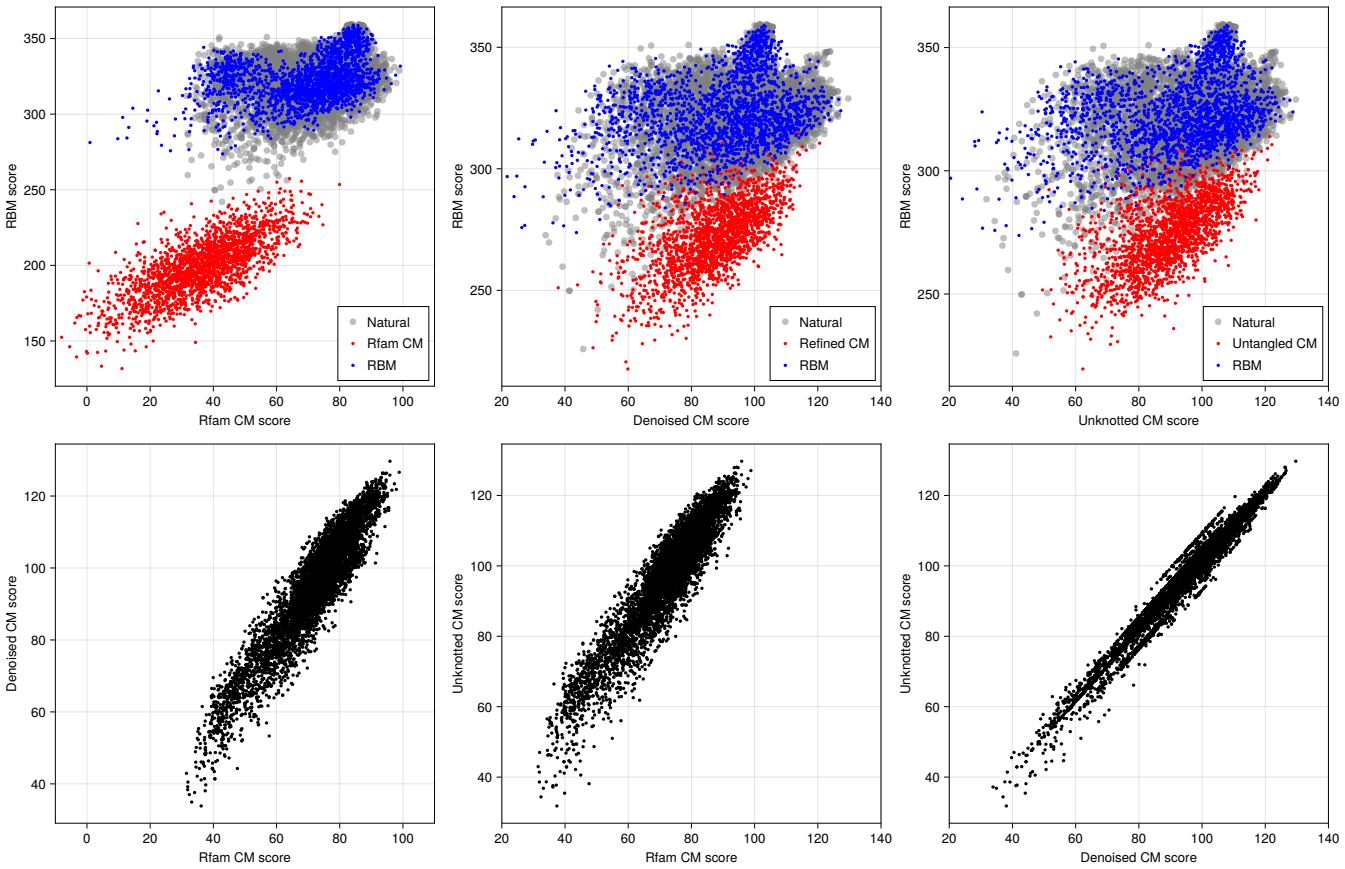

FIG. S11. In the panels in the top row, the x-axis gives CM scores and the y-axis gives RBM scores, for different classes of sequences. The bottom panels compare CM scores under three CM models of natural sequences. **(A)** Rfam CM. **(B)** Denoised CM. **(C)** Unknotted CM. **(D)** Rfam CM scores vs. Denoised CM scores, for natural sequences. Pearson correlation 0.93. **(E)** Rfam CM scores vs. Unknotted CM scores, for natural sequences. Pearson correlation 0.93. **(F)** Denoised CM scores vs. Unknotted CM scores, for natural sequences. Pearson correlation 0.98.

Appendix F: Distance of RBM generated aptamers to closest natural sequences

The RBM is capable of sampling novel aptamer sequences, distinct from any natural sequence. An example of this is shown below, in Fig. S16E, which shows that typical RBM samples are over 20 mutations away from the closest natural sequence. Figure S12 plots the Hamming distance to the closest natural sequence (x-axis) vs. the RBM effective energy $E_{\text{eff}}$of the probed RBM generated sequences in the first experiment. We find functional sequences up to 30 mutations away from the closest natural sequence.

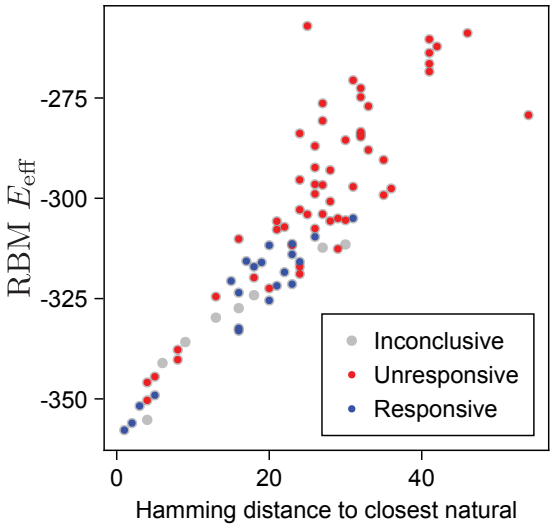

FIG. S12. Distance to closest natural sequence of RBM generated probed sequences. Blue points correspond to sequences that respond to SAM, according to the definition in Figure **??** of the main text. Red points denote sequences that do not respond to SAM, and gray points denote probed sequences for which the experimental observation is not conclusive about SAM response.

Appendix G: Replicate experiments

An independent experimental replicate was performed where the same set of 306 sequences was probed by SHAPE-MaP (following the same protocol as in the main text, Section IV G). We then applied the same analysis pipeline described in the main text (see Section IV H of the main text) to the replicate experimental data. The overall response to SAM of functional sequences was weaker in both natural and artificial aptamers. Although 112 of the probed aptamers were responsive in the first experiment (hereby called Replicate 0, see Fig. 6), the total number of responsive sequences in the replicate (hereby called Replicate 1) was 75 (see Table S13). Among the 75 functional aptamers in Replicate 1, 60 (80%) were also responsive in the first experiment, suggesting that our analysis pipeline is capable of delivering statistically consistent conclusions, across independent experimental realizations.

Fig. 6 and Table S13 summarizes these results. Figure S14 compares the reactivities in the replicates to each other in the different conditions probed.

A)

Group	Conclusive	Switchers	Non-switchers
Natural	131 of 201	70 (53.4 ± 4.4%)	61 (46.6 ± 4.4%)
Nat.(Seed)	97 of 151	56 (57.7 ± 5.0%)	41 (42.3 ± 5.0%)
Nat.(Hits)	34 of 50	14 (41.2 ± 8.4%)	20 (58.8 ± 8.4%)
Nat.(RBMscore>300)	87 of 137	47 (54.0 ± 5.3%)	40 (46.0 ± 5.3%)
Nat.(RBMscore>310)	62 of 96	38 (61.3 ± 6.2%)	24 (38.7 ± 6.2%)
Rfam CM	15 of 16	0 (0%)	15 (100%)
RBM	50 of 84	7 (14.0 ± 4.9%)	43 (86.0 ± 4.9%)
RBM(RBMscore>300)	33 of 53	7 (21.2 ± 7.1%)	26 (78.8 ± 7.1%)
RBM(RBMscore>310)	26 of 40	7 (26.9 ± 8.7%)	19 (74.1 ± 8.7%)
All	196 of 301	77 (39.3 ± 3.5%)	119 (60.7 ± 3.5%)
All(RBM score>300)	120 of 190	54 (45.0 ± 4.5%)	66 (55.0 ± 4.5%)
All(RBM score>310)	88 of 136	45 (51.1 ± 5.3%)	43 (48.9 ± 5.3%)

B)

	Responsive	Non-responsive	Inconclusive	**Total**
Switcher	**68**	2	0	70
Non-switcher	38	**23**	0	61
Inconclusive	39	4	**27**	70
Total	145	29	27	**201**

FIG. S13. Like Fig. 6C,F of the main text, but for the Replicate experiment.

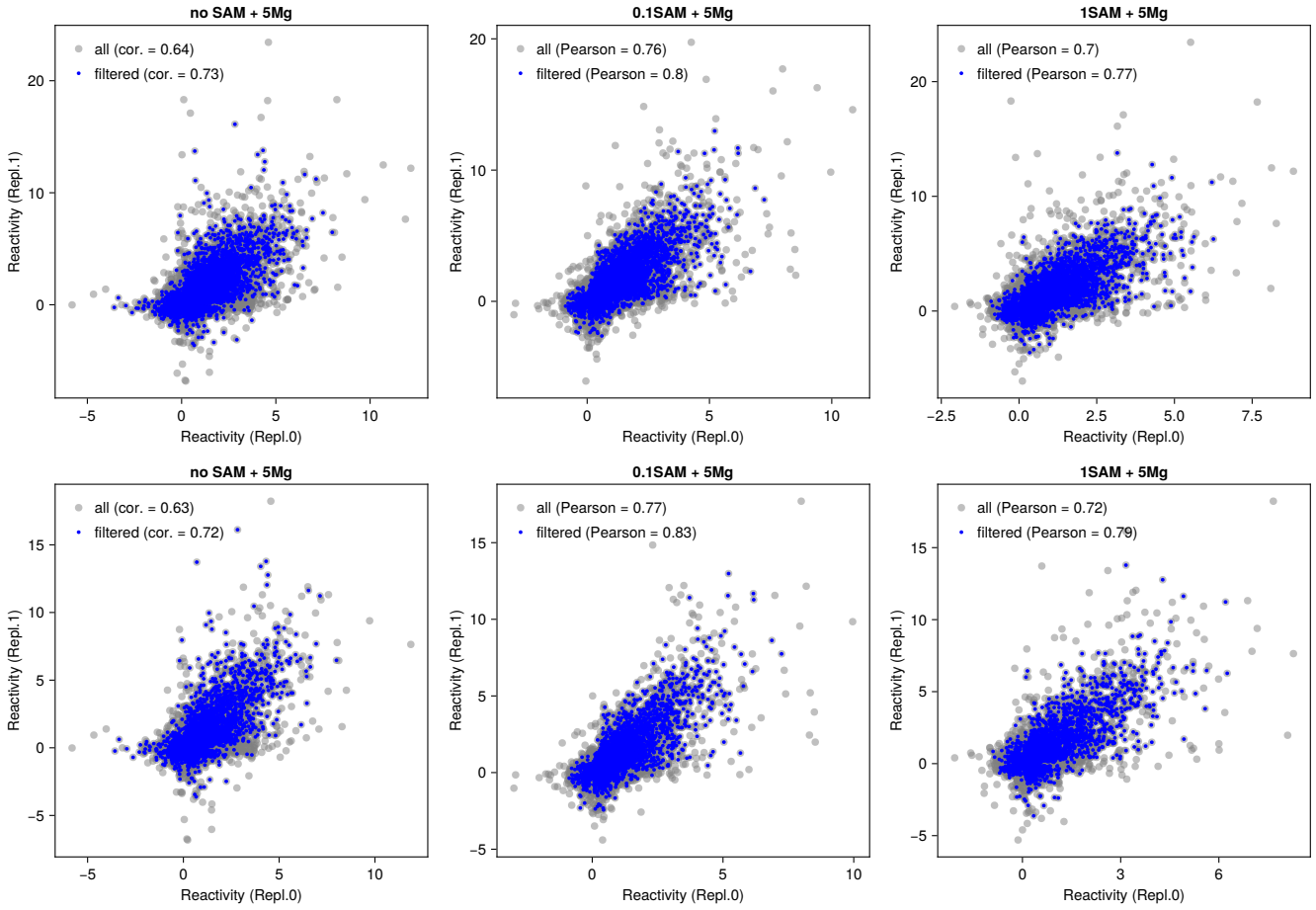

FIG. S14. Comparison of reactivities in both replicates (for the first experimental run). Each panel corresponds to one experimental condition. Pearson correlations are indicated in the legend, for all reactivities (gray), and for filtered reactivities after removing the most noisy values. The retained blue points satisfy: $\mathrm{stderr}(R) < 0.5\sqrt{|R|}$ for both replicates, where R is the SHAPE reactivity and $\mathrm{stderr}(R)$ the standard error, as reported by ShapeMapper [8]. The top row plots all reactivities, while the bottom row is restricted to natural aptamers only (excluding all artificial ones).

FIG. S15. Comparison of reactivities in both replicates at two concentrations of SAM.

Appendix H: RBM reproduces statistics of natural sequences

The quality of the model fit after training, and the quality of convergence, can be assessed by comparing statistics of sampled sequences against the empirical statistics of the MSA. In Figure S16A, we compare the single-site statistics, computed as the frequency of occurrences of each nucleotide (or gap symbol) at each position of the alignment. The agreement is excellent (Pearson correlation = 0.98), indicating that the RBM reproduces the conservation of important sites (cf. Figure 1C). Furthermore, RBM sampled sequences reproduce the covariance of pairs of sites of the natural sequences, as shown in Figure S16B. In this case, we compute the deviation of the frequencies of co-occurring nucleotides at pairs of sites from the expectation arising from their independent conservations. Such joint covariations arise from interactions across the sequence, related for example to secondary or tertiary contacts, or other functional constraints. The agreement is also excellent (Pearson correlation = 0.97), indicating that the RBM is able to reproduce the covariation of the natural MSA. We also evaluated the RBM effective energies E_{eff} (5) of sampled sequences compared to the energies assigned by the RBM to the natural sequences. As we show in Figure S16C, the two histograms are in close agreement to each other.

To evaluate the diversity of a set of sequences, natural or generated, we compute the matrix of all possible pairwise Hamming distances between pairs of distinct sequences, where the Hamming distance is defined as the number of positions where the two sequences differ. Figure S16D shows the histogram of these pairwise distances for the natural sequences in gray. Typically, two randomly selected natural sequences differ in about 40% of sites, or 43 out of the 108 aligned sites. We then sampled 5000 sequences from the RBM, and computed the histogram of their pairwise distances (between themselves). We plot the result in Figure S16D in red. We see that the histogram closely resembles the histogram of the natural sequences. We conclude that the RBM generated sequences recapitulate the natural diversity of the sequence homologs family. Furthermore, the RBM generates novel sequences, not seen in the data. Indeed, Figure S16E shows the histogram of distances between each sampled sequence, and the closest natural sequence to it. Typical RBM samples differ in 20 sites from the closest sequence in the MSA, and therefore constitutes a truly novel sequence.

Finally, we observed a strong variation in sequence lengths in natural sequences. In particular, dramatic variations of the P4 helix have been reported in the literature [7, 9], where riboswitches without P4 have been shown to be functional although with lower affinities to SAM. Although our RBM is not able to model insertions, it is still able to emit sequences of varying lengths by having more or less gaps in the sequence. We therefore compared the distribution of sequence lengths generated by the RBM, with the histogram of natural sequence lengths (not considering inserts) from the MSA. The plot in Figure S16F confirms that the RBM reproduces the correct length statistics.

Overall, these results suggest that the RBM is able to reproduce accurately several statistical features of the natural sequences.

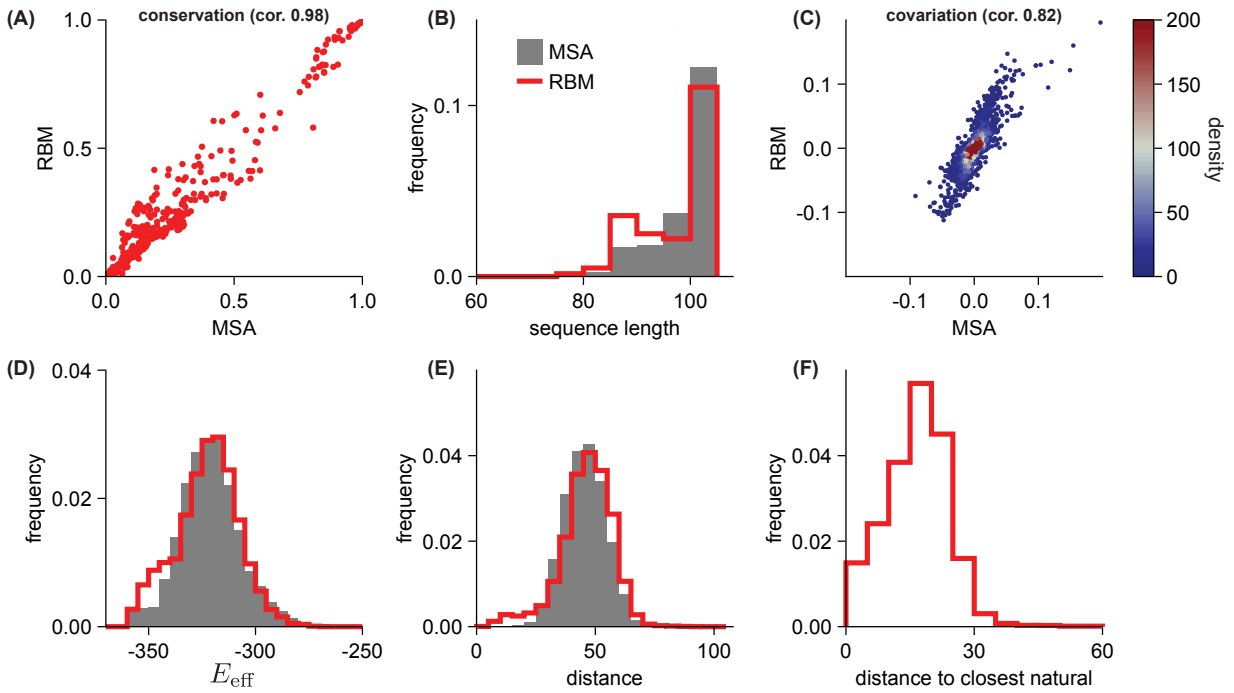

FIG. S16. RBM generates novel and diverse sequences that recapitulate statistics of natural homologues. **(A)** RBM samples reproduce single-site nucleotidic conservation of the natural sequences (Pearson correlation = 0.98). **(B)** Histogram of natural sequence lengths (gray) and of RBM generated sequences (red). Note that insertions are discarded. Sequence length is defined as the number of aligned sites that are not gaps (deletions). **(C)** RBM samples reproduce the statistics of nucleotidic covariation of natural sequences (Pearson correlation = 0.82). Since the number of paired sites is very large, the points are colored by their density in the plot, according to the colorbar legend. **(D)** Histograms of effective energies E_{eff} (Eq. (5) in the main text) of natural sequences (gray) and of RBM samples (red). **(E)** Histograms of pairwise Hamming distances, among natural sequences (gray) and among RBM samples (red). **(F)** Histogram of Hamming distances, from each RBM sampled sequence, to its closest natural sequence.

146

Appendix I: SHAPE analysis

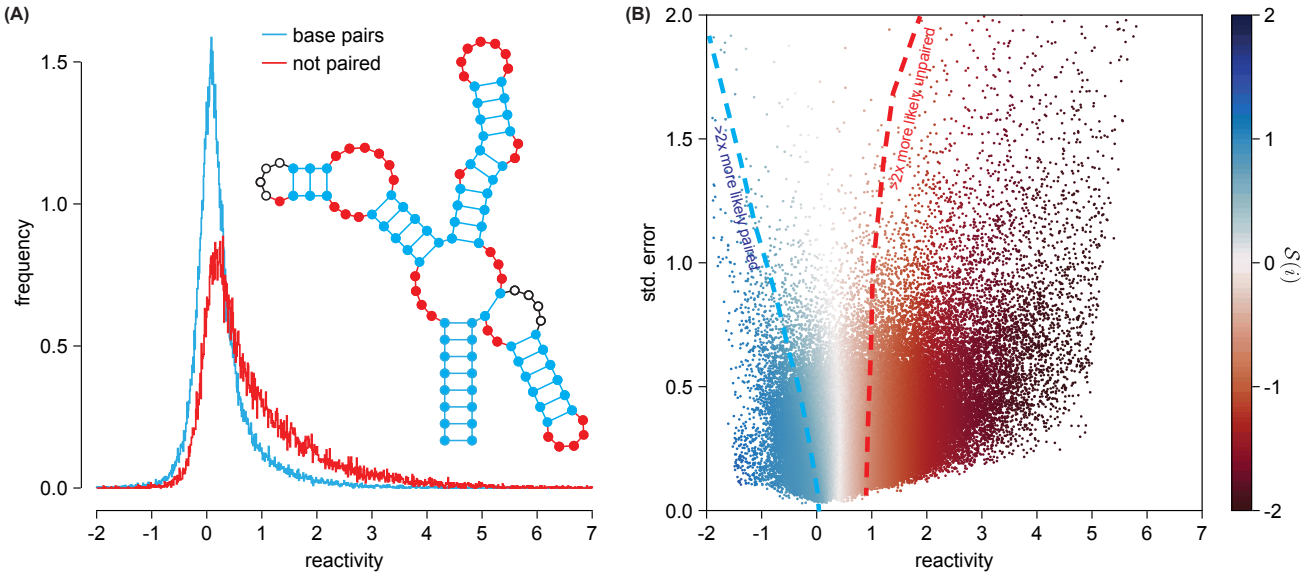

FIG. S17. Statistical differences of SHAPE reactivities in paired and unpaired sites. **(A)** Histogram of reactivities of base-paired (blue) and unpaired sites (red) in probed natural sequences belonging to the manually curated seed alignment. The inset shows the consensus secondary structure with sites colored according to whether they are paired or unpaired. Sites with ambigous behavior (pseudoknot, indicated by empty circles in the inset secondary structure) are excluded from both histograms. **(B)** Scatter plot of measured reactivities (on the x-axis) and their estimated standard errors (on the y-axis), as estimated by the standard ShapeMapper protocol [8], by first-order error propagation through the Poisson statistics of the mutation counts. The points are colored by the value of the log-odds ratio $\ln(P(\tilde{r}|\text{bp})/P(\tilde{r}|\text{np}))$ (see color bar), computed as explained in the text surrounding equation (15). The blue (red) dashed line indicates a contour separating sites over two times more likely to be paired (unpaired) than not.

Appendix J: Flexibility of the aptamer

Some of the RBM generated sequences classified as non-switchers, exhibit nonetheless localized reactivity responses in response to SAM, compatible with SAM binding, but not enough to conclude that a structural switch from the open to a closed conformation has taken place. Focusing on the first experiment, we find a total of 8 non-switcher RBM sequences exhibiting localized reactivity responses:

- 2 that respond at the SAM binding pocket
- 3 at the pseudoknot
- 3 at A minor
- 1 at base-triple

These are all different sequences. These results again highlight the diversity of structural responses displayed by the aptamer sequences tested in our experiments.

[1] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[2] T. Tieleman, Training restricted boltzmann machines using approximations to the likelihood gradient, in *Proceedings of the 25th international conference on Machine learning* (2008) pp. 1064–1071.
[3] J. Melchior, A. Fischer, and L. Wiskott, How to center deep boltzmann machines, The Journal of Machine Learning Research **17**, 3387 (2016).
[4] J. Fernandez-de Cossio-Diaz, S. Cocco, and R. Monasson, Disentangling representations in restricted boltzmann machines without adversaries, Physical Review X **13**, 021003 (2023).
[5] G. E. Hinton, A practical guide to training restricted boltzmann machines, in *Neural networks: Tricks of the trade* (Springer, 2012) pp. 599–619.
[6] J. Tubiana, S. Cocco, and R. Monasson, Learning protein constitutive motifs from sequence data, Elife **8**, e39397 (2019).
[7] B. Heppell, S. Blouin, A.-M. Dussault, J. Mulhbacher, E. Ennifar, J. C. Penedo, and D. A. Lafontaine, Molecular insights into the ligand-controlled organization of the sam-i riboswitch, Nature chemical biology **7**, 384 (2011).
[8] N. A. Siegfried, S. Busan, G. M. Rice, J. A. Nelson, and K. M. Weeks, Rna motif discovery by shape and mutational profiling (shape-map), Nature methods **11**, 959 (2014).
[9] J. J. Trausch, Z. Xu, A. L. Edwards, F. E. Reyes, P. E. Ross, R. Knight, and R. T. Batey, Structural basis for diversity in the sam clan of riboswitches, Proceedings of the National Academy of Sciences **111**, 6624 (2014).