Designing Molecular RNA Switches with Restricted Boltzmann Machines

Jorge Fernandez-de-Cossio-Diaz [a,1], Pierre Hardouin [a,2], Francois-Xavier Lyonnet du Moutier [2], Andrea Di Gioacchino [1], Bertrand Marchand[3], Yann Ponty[3], Bruno Sargueil[b,2], Rémi Monasson [b,1], Simona Cocco [b,1]

[a] *Equal contribution;* [b] *Corresponding authors;* [1] *CNRS UMR 8023, Laboratory of Physics of the Ecole Normale Supérieure & PSL Research, Sorbonne Université, 24 rue Lhomond, 75005 Paris, France;* [2] *CNRS UMR 8038, CitCoM, Université de Paris, 4 avenue de l'observatoire, 75006 Paris, France* [3] *CNRS UMR 7161, LIX, Ecole Polytechnique, Institut Polytechnique de Paris, 1 rue Estienne d'Orves, 91120 Palaiseau, France*

Riboswitches are structured allosteric RNA molecules that change conformation in response to a metabolite binding event, eventually triggering a regulatory response. Computational modelling of the structure of these molecules is complicated by a complex network of tertiary contacts, stabilized by the presence of their cognate metabolite. In this work, we focus on the aptamer domain of SAM-I riboswitches and show that Restricted Boltzmann machines (RBM), an unsupervised machine learning architecture, can capture intricate sequence dependencies induced by secondary and tertiary structure, as well as a switching mechanism between open and closed conformations. The RBM model is then used for the design of artificial allosteric SAM-I aptamers. To experimentally validate the functionality of the designed sequences, we resort to chemical probing (SHAPE-MaP), and develop a tailored analysis pipeline adequate for high-throughput tests of diverse homologous sequences. We probed a total of 476 RBM designed sequences in two experiments, showing between 20% and 40% divergence from any natural sequence, obtaining $\approx$ 30% success rate of correctly structured aptamers that undergo a structural switch in response to SAM.

I. INTRODUCTION

Riboswitches are regulatory RNA elements found mostly in bacterial and in some eukaryotic messenger RNAs. Usually located upstream of coding sequences, they modulate the expression of the downstream gene at the transcriptional or translation level in the presence of a specific metabolite [44, 61, 71, 73]; some riboswitches placed within genes even regulate alternative splicing [37]. In order to perform their function, these RNA motifs switch between two stable conformations in response to binding of their cognate metabolite to the aptamer domain of the riboswitch (Figure 1). This change of conformation, in turn, affects the expression platform, where the regulation signals are located. Understanding how the aptamer domain by itself is able to implement a structural switch in response to the ligand, and how this is encoded in the sequence, is an important step towards understanding the full riboswitch regulation.

The sequence-to-function mapping of structured RNAs is a complex problem. In the course of evolution, sequence patterns necessary for function are conserved, suggesting that large sequence datasets can shed light on this mapping. Comparative analysis of homologous RNA sequences collected in Multiple Sequence Alignments (MSA) [52] have been applied to predict RNA structures at the secondary or tertiary levels [8, 22, 28, 48, 59]. Covariation analysis has also been used to predict pseudoknots and other tertiary contacts from statistical couplings inferred from conservation and covariation across the MSA columns [15, 86], or by including positive and negative evolutionary information such as in the Cascade covariation Folding Algorithm (CacoFold) [58]. Machine learning approaches have recently shown promising results in RNA structure prediction. Among them Rosetta FARFAR2 [83] uses Monte-Carlo-based fragment assembly methods and can be aided by geometric deep learning approach such as ARES [78] to score putative structures. DeepFoldRNA [55] significantly outperformed the state-of-the-art tertiary structure prediction from sequence only. Although these approaches look promising, it is still an open question to understand why AlphaFold-level accuracies [34] are not reached in RNA structure prediction [57, 75].

The mirroring problem of designing RNA sequences capable of folding in a particular target structure or of performing a desired function has also long been investigated. Much effort has been devoted to the rational design of secondary structures, in particular with minimum free energy (MFE) approaches [25, 87]. Such approaches, due to computational limitations [59, 82], often ignore pseudoknots and other tertiary contacts, which are however known to be biologically important. To date, building generative models effective in designing RNA sequences with tertiary structural targets remains a challenging problem. From this point of view, riboswitches, besides their fundamental interest in biology and relevance for the RNA world hypothesis [37], offer a very interesting and difficult design problem, as their sequences encode not only two conformational structures but also a metabolite-mediated switching mechanism between them. In this paper, we address this challenging issue and show how to design functional RNA switches.

One of the largest groups of identified riboswitches recognize S-adenosyl methionine (SAM) as their effector metabolite [27, 56]. While six different SAM binding

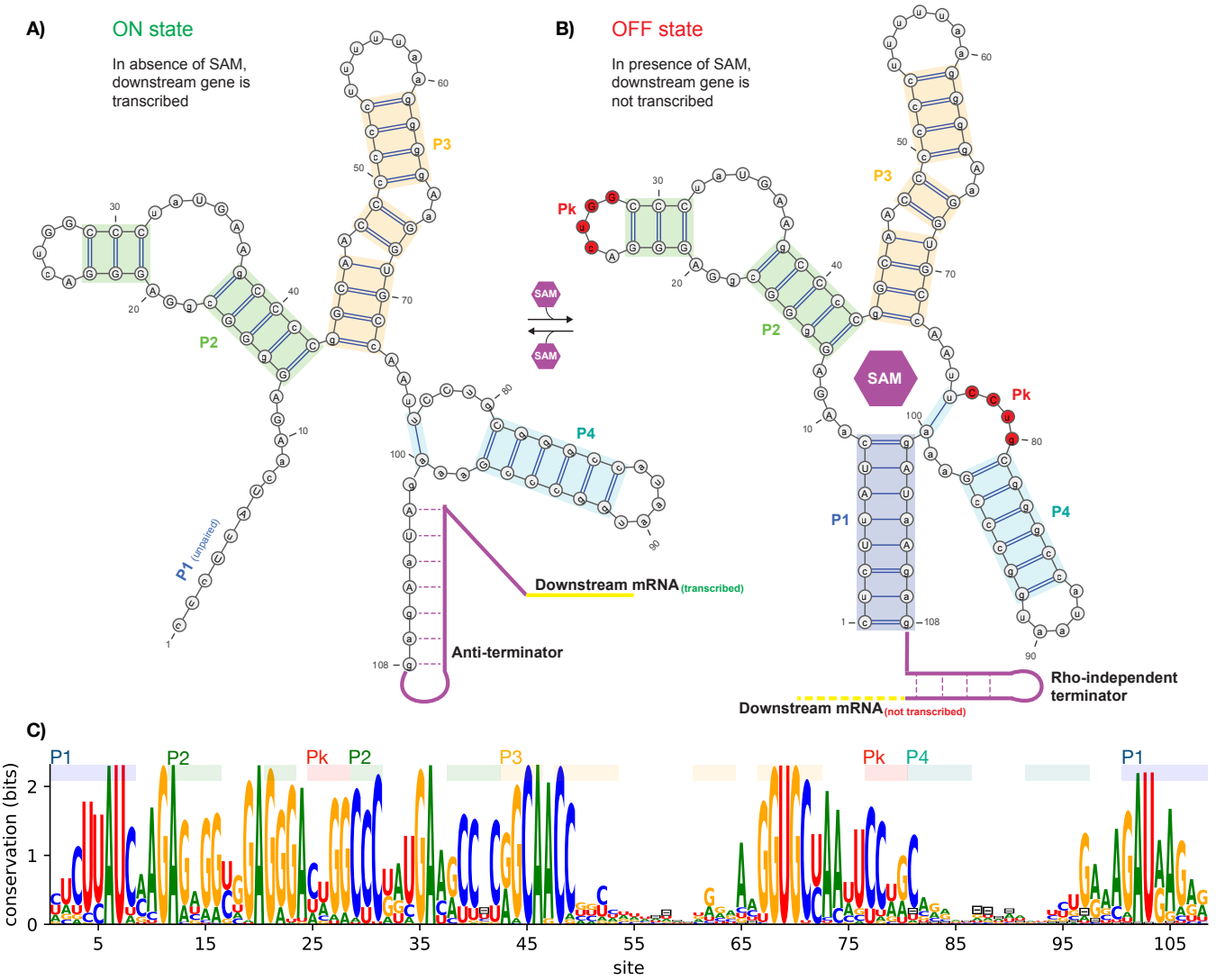


FIG. 1. Structure, regulatory function, and sequence conservation of the aptamer domain of the SAM-I riboswitch, acting at a transcriptional level. **A)** In absence of SAM, the P1 helix of the aptamer domain is unpaired, leaving the 3'-end free to pair with the anti-terminator segment of the expression platform. This conformation is incompatible with formation of a terminator motif, resulting in transcription of the downstream gene (ON state). **B)** SAM (represented by the purple hexagon) is captured in a groove contacting several sites around the central four-way junction. In the bound-state conformation, the P1 helix is fully base-paired. The expression platform is then free to form a Rho-independent terminator hairpin, which stops transcription of the nascent RNA, thus blocking the expression of a downstream gene (OFF state). The figure also shows several structural elements of the consensus secondary structure of the aptamer domain, including helices P1, P2, P3, P4, and a pseudoknot (Pk) in red. Secondary structure plots are obtained with VARNA [13]. **C)** Sequence conservation logo of aligned homologs of the SAM-I riboswitch aptamer domain family (RF00162 on Rfam). Gaps are indicated by the character '⊟'.

structural motifs have been identified, this study focuses on those harbouring a type I SAM aptamers (SAM-I) [1]. Figure 1A shows the secondary structure of the aptamer domain in absence of SAM, where transcription is allowed (ON state), while panel B depicts the structure when SAM is bound and transcription continuation is prevented (OFF state). Upon SAM binding, the aptamer cooperatively folds into the closed structure characterised by the stabilisation of P1, three triple base pairs and a pseudoknot (red in the figure) [62]. The closed state of the aptamer is stabilized by direct tertiary contacts between SAM and specific nucleotides forming the SAM binding pocket [49, 56]. To design new functional SAM-mediated RNA switches, we employed Restricted Boltzmann machines (RBM), a two-layer generative neural network, see Fig. 2, which can be regarded as building blocks of deeper neural architectures [18, 24, 31]. RBMs have been successfully applied to di-

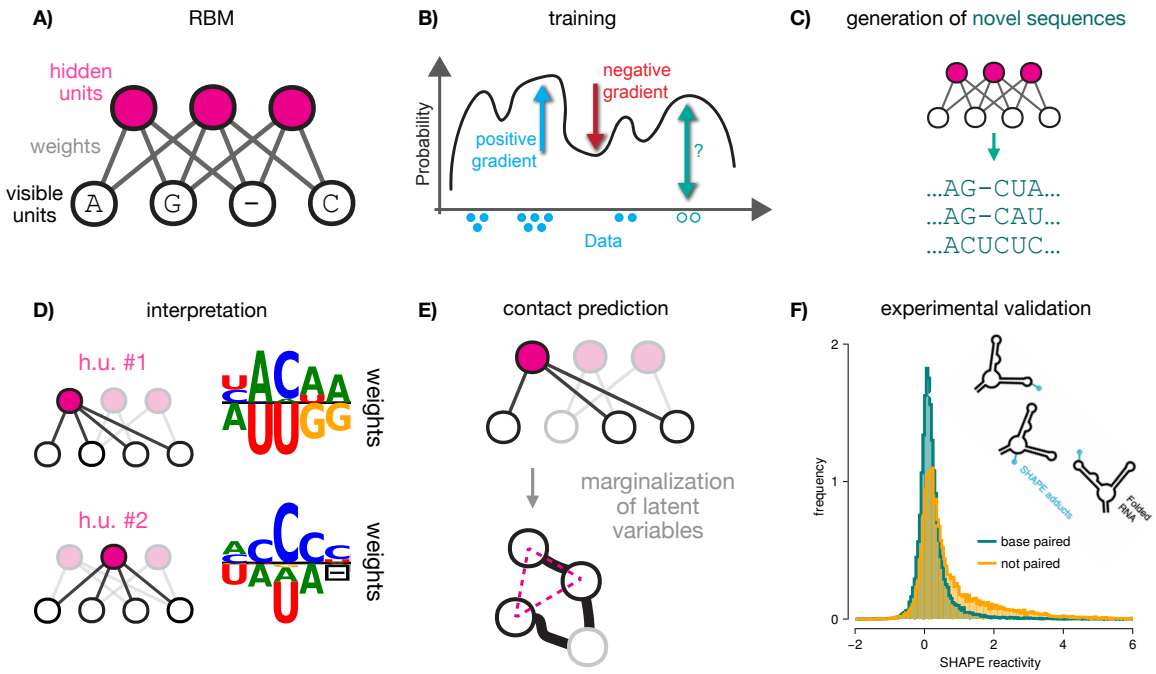


FIG. 2. RNA generative modelling with RBM and experimental validation. **A)** A Restricted Boltzmann machines (RBM) consists of two layers: visible and hidden. Each visible unit represents a site in an aligned sequence, and takes values A, C, G, U, or – (the alignment gap symbol). The two layers are connected by *weights*. No connections are allowed within units of the same layer. **B)** The RBM is trained by maximization of a regularized likelihood, see Eq. (9). A gradient term moves probability mass towards regions in sequence space populated by data, automatically discovering features desirable for functional sequences (blue). An opposite gradient term removes probability mass from regions void of data, automatically discovering constraints that, if violated, result in non-functional sequences (red). After training, the RBM may also assign large probability to potentially interesting sequences not covered by data (teal). **C)** The model can be sampled to generate novel sequences (teal), that often significantly differ from the natural ones. **D)** Hidden units extract latent features (nucleic-acid motifs) through the weights; weight values can be either positive or negative, and are shown above or below the zero weight horizontal bars in the logo plots, see Methods. Combining these motifs together allows RBM to design functional RNA sequences. **E)** The RBM is able to model complex interactions along the RNA sequence. Here, a hidden unit interacting with three visible units is highlighted. After marginalizing over the hidden-unit configurations, effective interactions arise between the visible sites, see Eq. (5). Here we represent schematically a three-body interaction, arising from the three weights of the marginalized hidden unit. **F)** Designed sequences are tested experimentally with SHAPE-MaP, a chemical probing approach. The differences in the reactivities of paired and unpaired nucleotides gives information about the structure of the RNA molecule [26, 64].

verse machine learning tasks [23, 65], and more recently to provide interpretable models of proteins in various contexts [4, 5, 47, 80], with application to design [19, 43]. We hereafter use RBM to develop a probabilistic model of sequences in the SAM-I riboswitch family. By learning those sequence statistics, the RBM models the constraints that enable aptamers to adopt the correct secondary structure, form tertiary contacts and moreover to effect a conformational switch in response to SAM.

II. RESULTS

Our pipeline is described in Fig. 2, and includes sequence data acquisition from Rfam [36], training of the RBM, sampling of the model to design new artificial SAM-I aptamers, and lastly experimental validation of SAM-induced conformational switch in designed sequences by SHAPE-MaP probing [14, 69].

A. Generative models of SAM riboswitches

We trained an RBM on a multiple sequence alignment (MSA) of natural homologues of the SAM-I riboswitch, gathered from the Rfam [36] database (Rfam ID: RF00162). RBM are energy-based generative models, that once trained, define a score, $-E_{\rm eff}(\mathbf{v})$, over all possible sequences $\mathbf{v}$. Sequences with high scores (equivalently, low energies) are then "good" fits to the family, according to the model. Novel sequences of high score can be generated by sampling the resulting Boltzmann measure, $P_{\rm RBM}(\mathbf{v}) \propto e^{-E_{\rm eff}(\mathbf{v})}$, see Methods for details. Although there are no direct couplings between visible

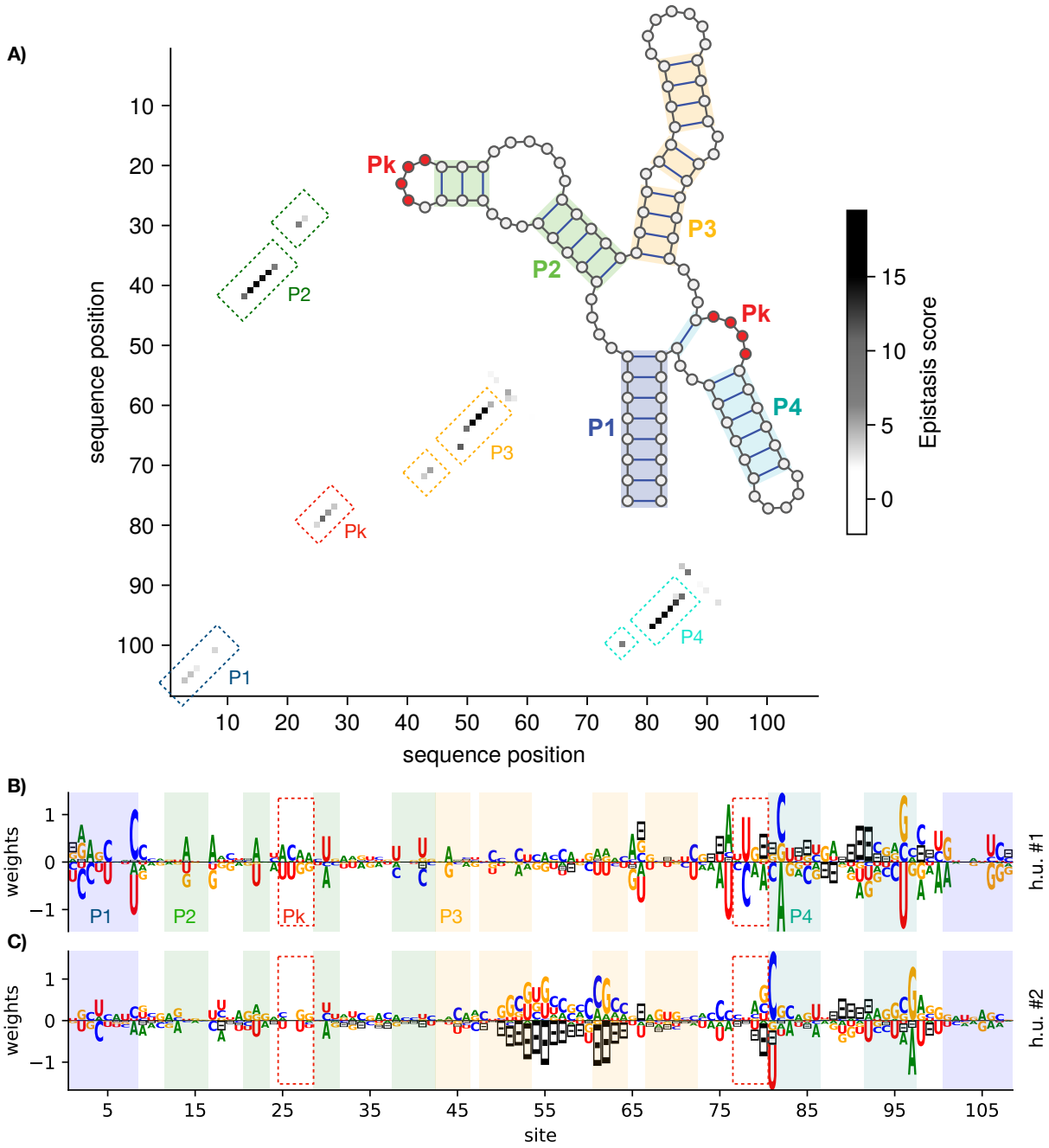


FIG. 3. Interpretation of RBM extracted features. **A)** Contact map, constructed by computing an epistatic score from the RBM marginal probability over sequences, see Eq. (11) and (12). The highest epistatic scores correspond to major secondary and tertiary contacts of the SAM-bound aptamer structure, shown in the inset. **B,C)** Sequence logos of the weights $w_{i\mu}(v_i)$ attached to exemplary hidden units (#1 and #2) of the RBM, selected by having the highest weight norms. Each letter size in the logo is proportional to the corresponding weight, see Figure 2D and [76, 80]). Sites are colored according to the secondary structure element they belong to, including the paired (P) helices P1 (light purple), P2 (green), P3 (yellow), and P4 (teal). Sites participating in the pseudoknot (Pk) are also highlighted (red dashed box). In hidden unit #1, Watson-Crick complementarity along P1 (*e.g.*, site 8 with 101) is favored, in agreement with base pairing of these positions at the 5' and 3' ends of the P1 helix. The same unit also puts weights on complementarity along the pseudoknot (*e.g.* sites 25-28 with 77-80), helping stabilize this tertiary contact. The fact that these complementarity constraints, belonging to different structural motifs, are enforced by the same unit, suggests that P1 and the pseudoknot stabilize in a concerted manner (*c.f.* Fig. 1) in response to SAM. Hidden unit #2, on the other hand, places significant weight in the complementarity between sites 81 and 97, stabilizing P4 and along various P3 sites, favouring a dichotomy between stabilizing complementarity or deletions in this segment. Indeed, some natural sequences lack a hairpin loop at P3 (sites 50–64), consistently with a negative activation of h.u. #2.

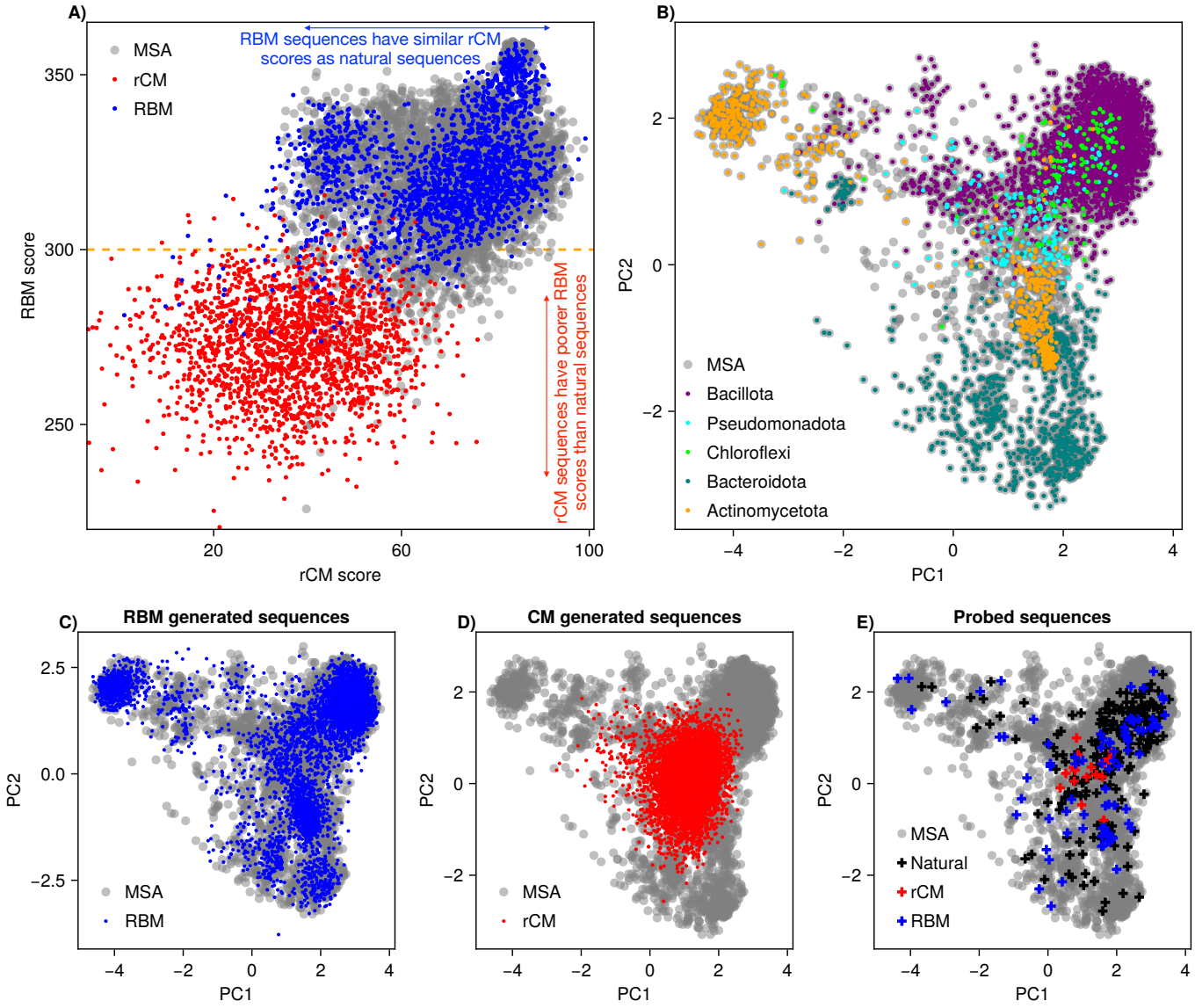


FIG. 4. Sequence generative models. **A)** Scatter plot of Rfam CM (rCM) scores (x-axis) vs. RBM scores (y-axis), for natural sequences (gray), rCM sampled sequences (red), and RBM sampled sequences (blue). A threshold at RBM score = 300 (orange dashed line) separates the majority of rCM generated sequences from the majority of RBM and natural sequences. **B)** Projection of natural MSA sequences (seed + hits) onto the top two principal components of the MSA correlation matrix (gray). The largest taxonomic groups (with > 100 member sequences) are highlighted in colors. Taxonomic annotations were obtained from NCBI. **C)** Projection of RBM generated sequences (in blue) on the top two principal components of the MSA, with the natural sequences in the background (gray). **D)** Projection of rCM generated sequences (in red) on the top two principal components of the MSA, with the natural sequences in the background (gray). **E)** Projection of all probed sequences on the top two principal components of the MSA, with the natural sequences shown in background (gray). The 301 probed sequences in the first experimental batch are colored by their origin: Natural (black), Rfam CM (rCM, red), and RBM (blue). The selection of probed sequences spans the wide diversity of natural sequences.

units in the RBM architecture (Fig. 2A), sites are coupled through the common links to the hidden units (Fig. 2E). The effective epistatic score between pairs of sites [80] (see Eq. (12) for precise definition), is plotted as a heatmap in Fig. 3A. Some pairs of sites have epistatic scores significantly larger than others, and correspond to contacts in folded aptamer, see mapping onto the secondary structure (Fig. 3A). Interestingly, epistatic scores at P1 are weaker than in other helices, reflecting the flexibility of P1, that is able to open or close in concert with SAM binding (*c.f.* Fig. 1). The pseudoknot is also correctly identified (red in Fig. 3A), proving the capability of RBM to identify tertiary motifs. Besides structural contacts, the RBM hidden units capture extended mo-

tifs, relevant for tertiary structure formation, see weights in Fig. 3B,C.

We then evaluate the quality of sequences designed by the RBM by comparing their scores to the ones of natural sequences and sequences designed by Covariance Models (CM). CM capture the conservation profile of important sites along the sequence, as well as correlations due to the complementarity of base-pairs in the secondary structure [21], but are unable to model tertiary motifs (such as pseudoknots). Rfam sequence alignments [35] are based on CM [52], which score and filter candidate member sequences of a given family. Our first baseline is the CM model underlying RF00162 (that we also call Rfam CM), that we downloaded directly from Rfam (Methods).

To assess how well the RBM models secondary structure constraints, we computed the Rfam CM (rCM) scores of sequences sampled from the RBM. In Fig. 4A, we show a scatter plot of rCM vs RBM scores for natural, RBM and CM generated sequences. RBM generated sequences have rCM scores comparable to the natural ones, indicating that RBM samples satisfy the constraints imposed by the rCM model, as well as the natural sequences. Moreover, RBM samples have RBM score comparable to natural ones while Rfam CM samples have scores significantly smaller (indicated by the orange threshold in the panel), suggesting that the RBM imposes further constraints beyond those imposed by the rCM, such as tertiary contacts (pseudoknot, A-minor motif), and possibly other functional constraints not representable by the CM, and which could be important for the function of the aptamer. In addition, RBM recapitulates several statistical properties of natural sequences in the MSA, including conservation, covariation, distribution of lengths, and others (see Supplementary Section H).

Next, we considered a principal component analysis (PCA) of the natural MSA. The top principal component (PC) captures a mode of variation associated to deletion of the P4 helix, as can be seen from the large number of gaps in this region (the two PCs are shown in Supplementary Fig. S5). Figure 4B shows the projections of the natural MSA sequences, annotated by their taxonomic class, onto the top two PCs. The most populated taxonomic classes are highlighted with colors. The PCs appreciably separate taxonomic clusters of natural sequences. In particular, a group of Actinomycetota, in the top left corner, have completely deleted or very short P4 helix segments. Literature reports suggest that SAM aptamers are able to function in absence of P4 [79], although their affinity towards SAM decays with decreasing P4 length [30].

We projected RBM generated sequences onto the two PCs, see Fig. 4C. The RBM samples span the full PCA space, covering all the clusters associated to different taxonomic classes. In contrast, Rfam CM generated sequences, shown in Fig. 4D, span a limited portion of the PC space, remaining confined to a central region in the plot. Since CM is able to model a limited amount of covariation, this suggests that the flexibility of RBM to better capture interactions in the data is important to model the full variability of the family of homologues.

To validate the generative capabilities of these models, we selected a subset of sequences and probed their ability to respond to SAM binding experimentally. Their PCA projections are shown in Fig. 4E, colored by their origin (Natural, Rfam CM, RBM). The selection spans a wide range of variability observed in the natural MSA. See Methods for details about the selected sequences.

B. SHAPE-MaP reactivities of natural and generated sequences

Our next step was to experimentally test the properties of the generated sequences. SHAPE-MaP is a high-throughput technique that reports structural information about RNA molecules [69]. In brief, SHAPE-Map measures the structural flexibility of individual nucleotides. Paired nucleotides are expected to show a low SHAPE reactivity, while those left single strand harbour a high shape reactivity. Similarly, in presence of SAM, positions that contact SAM are expected to be less reactive.

We sampled and probed in presence or absence of SAM a set of 208 natural sequences representative of Rfam ID RF00162 (see methods) shown by black crosses in Fig. 4E. After standard processing [69], SHAPE-MaP results in reactivity values $r_{i,n,c}$ assigned to each site i, for each aptamer n, and in each condition tested c. Figure 5A shows the histogram of reactivities of expected base-paired sites (teal) and unpaired sites (gold) in presence of SAM, colored according to the consensus secondary structure of the SAM-I riboswitch aptamer domain family; excluding sites involved in the pseudoknot. As expected, base-paired sites tend to have lower reactivities, and unpaired sites tend to have higher reactivities. This also confirms that most natural aptamers are compatible with the expected secondary structure [1].

Next, we considered the effect of SAM on the aptamer structure. Figure 5B (left) shows that the histogram of reactivities of the sites associated to the pseudoknot (black) in absence of SAM is compatible with the histogram of unpaired sites, consistently with the expected conformation of the aptamers in this condition (Fig. 1). In presence of SAM, the histogram of pseudoknot reactivities in Fig. 5B (right) agrees with the distribution of paired sites. This confirms that, upon SAM addition, a conformational switch occurs leading to formation of the pseudoknot. The same observation can be made for the P1 helix (Supplementary Fig. S4).

The difference in reactivities with and without SAM, $\Delta r_{i,n} = r_{i,n,\mathrm{SAM}} - r_{i,n,\mathrm{no\ SAM}}$, once averaged over all probed natural sequences n to remove experimental noise [69], defines a site-dependent Δ-reactivity template, $\langle \Delta r_i \rangle_{\mathrm{nat.}}$, shown in Figure 5C (in gray). We observe reactivity decreases (also called protection) for the pseudoknot (sites 25-28, 77, 79, 80), the base triple (24, 76, 100),

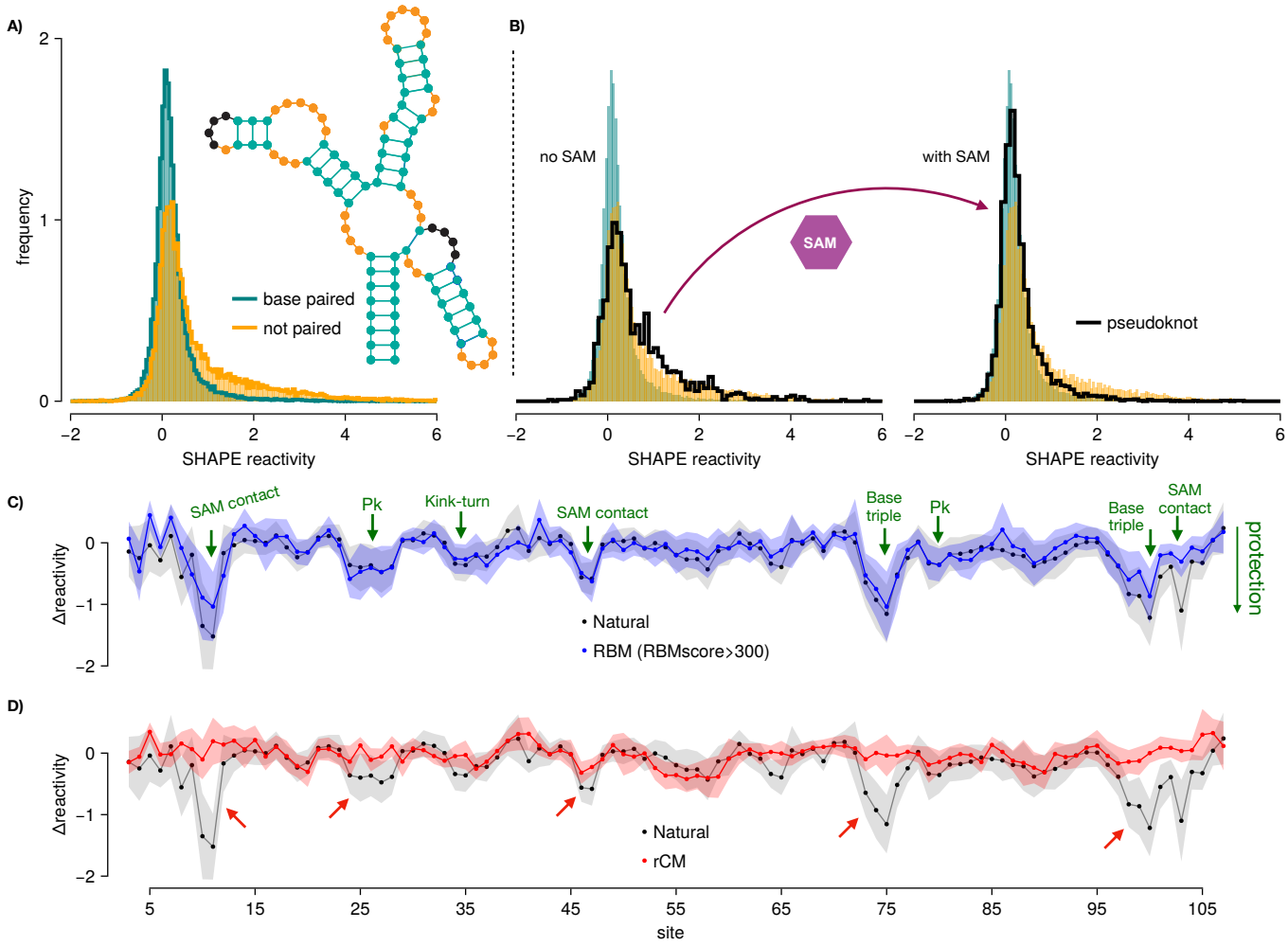


FIG. 5. SHAPE-MaP probing of aptamers. **A)** Histogram of SHAPE reactivities of base-paired (teal) and unpaired sites (gold) for the probed natural sequences. The inset shows the consensus secondary structure of the SAM-I riboswitch aptamer domain, highlighting base-paired (teal) and unpaired (gold) sites. The sites forming the pseudoknot in presence of SAM (black in the inset) are not included in these histograms. **B)** Histogram of SHAPE reactivities of the pseudoknot sites (black) in absence of SAM (left), and in presence of SAM (right). Without SAM, the distribution of reactivities of the pseudoknot sites agrees with the reactivities of unpaired sites. When SAM is added, the SHAPE reactivities of the pseudoknot sites move towards the distribution of paired sites. **C)** Average differential reactivities in response to SAM of RBM generated sequences with high RBM scores (> 300) (blue), across the 108 sites of the alignment. For comparison, the average differential reactivities for natural sequences are shown in the background (gray). High-RBM score sequences recapitulate protection of sites involved in the structural switch in response to SAM binding (highlighted in green). **D** Average differential reactivities in response to SAM of Rfam CM (rCM) generated sequences (red). Natural sequences are shown in background for comparison. Rfam CM sequences fail to recapitulate the expected protections associated to the structural switch (red arrows). In both panels (B,C), the bands indicate $\pm$ half one standard deviation.

for some of the sites directly in contact with SAM (10, 11, 46, 103) or forming the A-minor tertiary structural motif (73, 74), This indicates that the natural sequences probed in our experiments are, mostly, SAM binders (as expected), and moreover recapitulate expected structural changes upon binding. We also confirmed that members of the seed alignment (a manually curated subset [36]) show the same average reactivity responses (see Supplementary Fig. S7). Figure 5C also shows bands indicating the standard deviation of the Δ-reactivities.

Comparing the Δ-reactivity template of natural aptamers to the Δ-reactivity profiles of sequences generated with RBM (and having high score > 300) and with Rfam CM allows us to assess the functionalities of these generated sequences. The average Δ-reactivity profile of RBM generated sequences (in blue in Fig. 5C) is in excellent agreement with the one of natural sequences, exhibiting protections at key sites compatible with SAM binding and the expected structural switch. This suggests that RBM designed sequences of high enough RBM scores

displays, in average, the same behavior as their natural counterparts. We can also confirm that the RBM sequences with scores > 300 reproduce the reactivity response to magnesium of natural sequences (see Supplementary Fig. S8). These results suggest that RBM can generate functional aptamers. In contrast, RBM sequences with lower scores (< 300) show larger discrepancies (Supplementary Fig. S9).

Lastly, we show in Fig. 5D the average differential reactivities of sequences sampled from Rfam CM (in red). This group of sequences shows appreciable lack of protection at key sites, such as 10-11 (SAM contact), 25-28 (pseudoknot), 73-76 (base triples), and 103 (SAM contact in P1). These sequences therefore show no or weak response to SAM, indicating poor binding and/or lack of the structural switch necessary for regulatory function.

In summary, these results suggest that most natural sequences tested bind SAM and respond according to the expected structural switch. RBM generated sequences of high enough scores have statistically the same structural response to SAM as natural sequences. In contrast, Rfam CM generated sequences, or RBM sequences of lower scores, have markedly different responses, incompatible with the natural function (see also Supp. Fig. S9). We remark that these statements are valid as an average over the sequences in each group, and does not exclude the possibility that some particular sequences fail to bind SAM or respond in a different manner.

C. Detailed structural responses of individual sequences

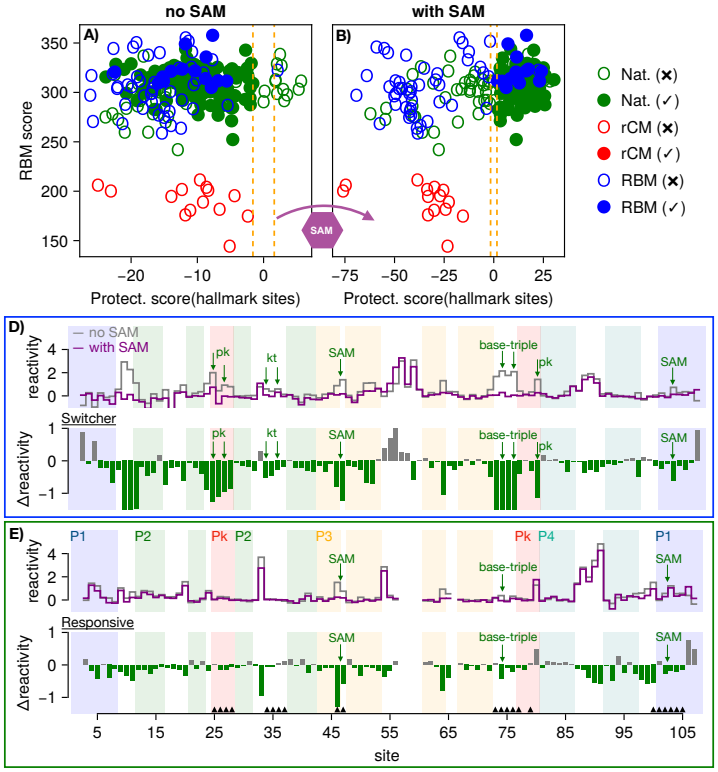
We investigated the response of individual sequences to SAM. Using the histograms of reactivities for paired and unpaired sites in Fig. 5A, we obtained a *Protection Score* $\mathcal{S}(\mathcal{M})$ that any subset of sites $\mathcal{M}$ be paired. The *Protection Score* is the log-likelihood ratio comparing the probability that the sites in $\mathcal{M}$ are paired and unpaired given their reactivities; it takes into account the sampling noise arising from the finite sequencing read depth [69], see Methods and [22]. We choose for $\mathcal{M}$ hallmark sites expected to show significant reactivity changes in functional aptamers in response to SAM binding: the pseudoknot, SAM contacts, and others previously reported in the literature [29] and highlighted by the global analysis of the natural aptamers Fig. 5D, shown in Fig 6 E and collected in Table I. Natural sequences exhibit robust responses at these sites in response to SAM, and their protection scores are compatible with fine structural details of the aptamers (see Extended Fig. 9).

We then analyzed all generated sequences. Figure 6 A and B plot $\mathcal{S}$(hallmark sites) without (left) and with (right) SAM, against their RBM scores. The structural switch in response to SAM is manifest on these hallmark sites: $\mathcal{S}$ shifts from negative values in the absence of SAM (indicating these sites are likely to be unpaired), to positive values in the presence of SAM (indicating that these sites are involved in an interaction). Sequences that switch from an open to a closed conformation in response to SAM, hereafter referred to as *switchers*), are shown as filled disks in the figure, while those that do not (*i.e.* either because the hallmark sites remain reactive or unreactive in both conditions) are shown with empty circles 6C. Aptamers that remain below the 5-fold significance threshold are listed as inconclusive.

We find that switcher sequences (both natural and designed) tend to have high RBM scores. In particular, 35% of RBM designed sequences above a score threshold of > 300 switch structurally in response to SAM, exhibiting significant response in the hallmark sites. These sequences have typical Hamming distances between 10 and 30 to the closest natural sequence, see Supplementary Fig. S12. There are two modes of failure for the remaining RBM sequences: either the structural motifs (pseudoknot, P1, etc.) are protected even in absence of SAM, or they remain reactive in presence of SAM. We find that most of the 45 RBM non-switcher sequences fail to do so in the second manner: they do not form the necessary contacts with or without SAM. On the other hand non-switcher natural sequences can fail in both modes. None of the sequences generated with Rfam CM are functional, possibly on account of the inability of the CM to model tertiary motifs like pseudoknots [21, 52].

Besides computing protection scores for all probed aptamers, we have also inspected manually the SHAPE Δ-reactivity profiles. Figure 6D shows an example of a RBM-generated switcher sequence for which the Δ-reactivity profiles are compatible with a global structural switch, as evidenced by reactivity changes (highlighted by arrows) in most of the hallmark sites I. In some cases, aptamers can exhibit localized reactivity changes in response to SAM, indicating ligand binding, but not enough to conclude a global structural switch from an open to a closed conformation has occurred. Fig. 6E shows an example of a responsive natural sequence for which the Δ-reactivity is localized on few of the hallmark sites; this strongly suggests a binding event but is not enough to conclude that a global structural rearrangement has occurred, hence this sequence is not identified as a switcher. Table 6F summarizes these analysis, where "responsive" denotes aptamers with at least some significant localized reactivity changes in response to SAM. Importantly, in comparison to the manual inspection for localized responses, the analysis based on protection scores is more stringent, tending to require global structural switch responses across most hallmark sites, and has the advantage of being automatic, objective and reproducible (we provide a source code package in SI).

We performed an experimental replicate on the same set of sequences. Although some aptamers in the replicate exhibited an overall lower response to SAM (natural and artificial), the fractions of switchers in each group are consistent with the results reported in Fig. 6. Moreover, 80% of identified switchers in the replicate are the same



C)	Conclusive	Switchers	Non-switchers
Natural	145 of 201	97 (66.9 ± 3.9%)	48 (33.1 ± 3.9%)
Nat.(Seed)	111 of 151	75 (67.6 ± 4.4%)	36 (32.4 ± 4.4%)
Nat.(Hits)	34 of 50	22 (64.7 ± 8.2%)	12 (35.3 ± 8.2%)
Nat.(RBMscore>300)	96 of 137	67 (69.8 ± 4.7%)	29 (30.2 ± 4.7%)
Nat.(RBMscore>310)	65 of 96	50 (76.9 ± 5.2%)	15 (23.1 ± 5.2%)
Rfam CM	14 of 16	0 (0%)	14 (100%)
RBM	59 of 84	14 (23.7 ± 5.5%)	45 (76.3 ± 5.5%)
RBM(RBMscore>300)	40 of 53	14 (35.0 ± 7.5%)	26 (65.0 ± 7.5%)
RBM(RBMscore>310)	31 of 40	12 (38.7 ± 8.7%)	19 (61.3 ± 8.7%)
All	218 of 301	111 (50.9 ± 3.4%)	107 (49.1 ± 3.4%)
All(RBM score>300)	136 of 190	81 (59.6 ± 4.2%)	55 (40.4 ± 4.2%)
All(RBM score>310)	96 of 136	62 (64.6 ± 4.9%)	34 (35.4 ± 4.9%)

F)	Responsive	Non-responsive	No data	**Total**
Switcher	**91**	5	1	97
Non-switcher	27	**19**	2	48
Inconclusive	8	2	**46**	56
Total	126	26	49	**201**
N-s./SAM pocket prot.	**7**	0	0	7
N-s./No SAM pocket prot.	20	**12**	0	32

FIG. 6. Detailed structural responses of individual sequences. **(A,B)** Protection scores $\mathcal{S}$ for the set of hallmark sites (Table I and Eq. (16)), vs. RBM scores for all probed sequences. Panels: (A) without SAM, (B) with SAM. Aptamers switching from the open conformation $\mathcal{S} < -\mathcal{S}_0$ without SAM to the closed conformation $\mathcal{S} > \mathcal{S}_0$ when SAM is present are shown with filled circles. We use a 5-fold significance threshold of $\mathcal{S}_0 = \ln(5)$ to discard inconclusive sequences (between the dashed orange vertical lines), see Methods. Colors refer to the sequence origin: Natural, Rfam CM (rCM), or RBM. **(C)** Numbers of switcher sequences in each group. Error bars reflect the uncertainty in the estimated fractions from the limited numbers of conclusive aptamers in each case (Methods). **(D)** Example reactivity profiles (without and with SAM) of a RBM generated aptamer (blue) identified as SAM switcher. Bottom panel shows Δ–reactivities in response to SAM. Background color correspond to consensus secondary structure. **(E)** Example reactivity profiles of a natural aptamer, classified by manual inspection as responsive. Reactivity responses (green arrows) are localized in some of the Hallmark sites corresponding to the SAM-binding pocket. The automated analysis considers this aptamer as non-switcher, but detects protection in the SAM binding pocket. The hallmark sites (Table I) are indicated by black triangles at the bottom. **(F)** Comparison of manual (columns) and automatic (rows) scoring of natural aptamers. The bottom two rows show how non-switchers are classified according to the protection scores of the SAM binding pocket sites.

switchers from the first experiment, confirming the robustness across experimental replicates of the protection score analyses. See Supplementary Table S13 for these results.

Inspired by previous experimental observations for other riboswitches [32, 84] and Sabatier's principle for enzymes, which require intermediate substrate binding energies for proper function [60], we computed the thermodynamic energies brought by P1 helix formation using the Turner energy model as implemented in the ViennaRNA package [40] (Methods). In order to be able to switch, molecules cannot have a P1 pairing energy too low, nor too high. Figure 7A shows that the sequences that respond to SAM through P1 helix stabilisation are confined to a thermodynamic energy window ranging from -10 to 0 kcal/mol (blue dashed lines). Similarly, the pseudoknot (Pk) formation in response to SAM tends to occur for aptamers having a Pk pairing energy between -8 and -3 kcal/mol (Fig. 7B). Since P1 consists of 8 base pairs, while the pseudoknot involves 4 base pairs, in both cases the flexible energetic window spans a range of 1.25 kcal/mol per base-pair, or about one third of the typical base-pairing energy [40]. The leftmost panels in Fig. 7 show that RBM samples preferentially have pairing energies in the expected intermediate band for both P1 and Pk, and are thus compatible with the structural switch required for riboswitch function.

The tables in Fig. 7 give a summary of these results. Interestingly, 25 out of the 27 aptamers that stabilize P1 in response to SAM are also switchers, in the sense of Fig. 6A,B, and show broad structural responses in other Hallmark sites (Table I). Similarly, 29 out of 32 aptamers

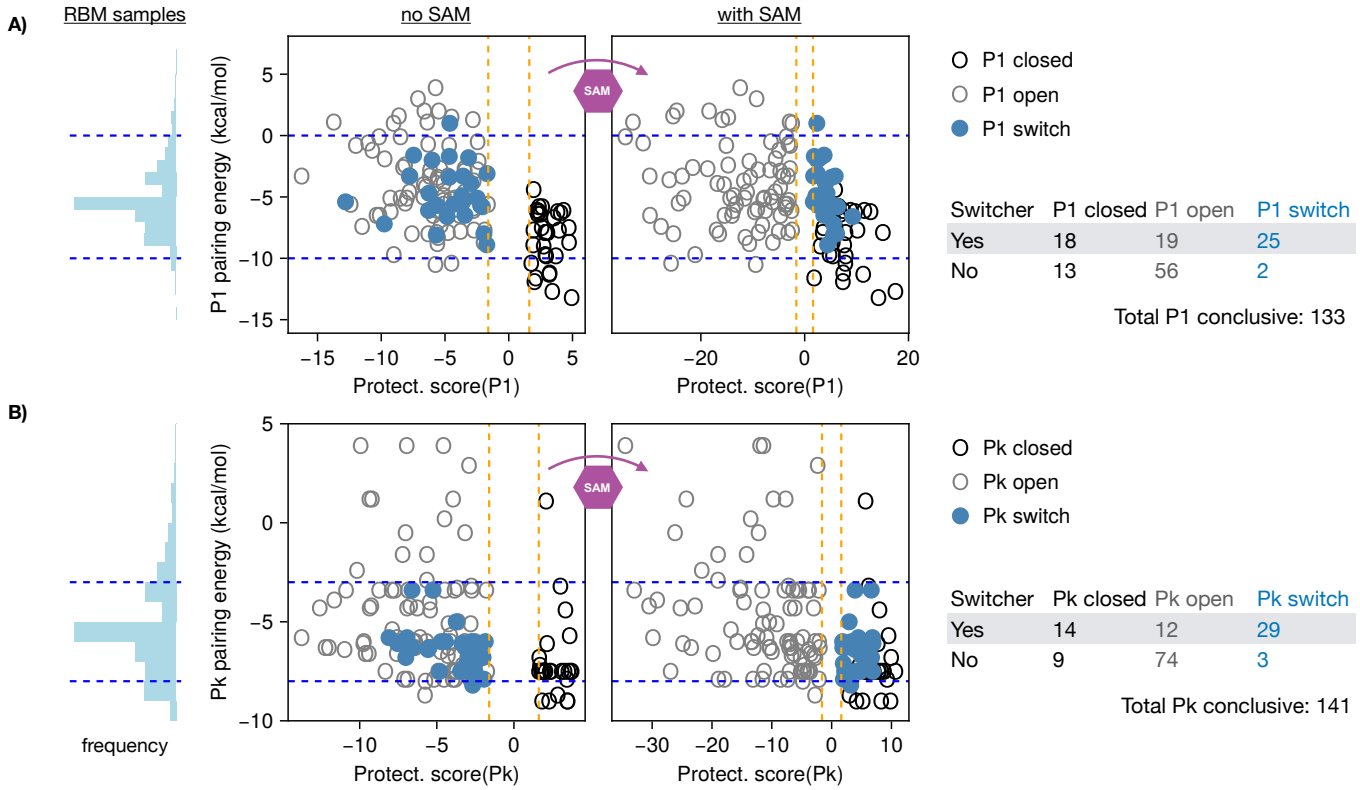


FIG. 7. Local switching responses along P1 and the pseudoknot (Pk) require intermediate pairing energies. **(A)** Left panel shows the histogram of Turner pairing energies for P1 (computed with the ViennaRNA package [40]) of RBM-designed sequences. The following panels show the protection scores $\mathcal{S}(P1)$ (x-axis) of P1 vs. pairing energy without and with SAM. Aptamers are colored according to their response: if $\mathcal{S}(P1) > \mathcal{S}_0$ in both conditions, P1 is always closed (open black circle); if $\mathcal{S}(P1) < -\mathcal{S}_0$ in both conditions, P1 is always open (open gray circle); if $\mathcal{S}(P1)$ crosses from one side to the other, the motif switches in response to SAM (filled light blue disks). The table lists then numbers of aptamers that show a structural switch in response to SAM in hallmark sites, compared to a local switch response in P1. **(B)** Like (A), but for pseudoknot (Pk) sites.

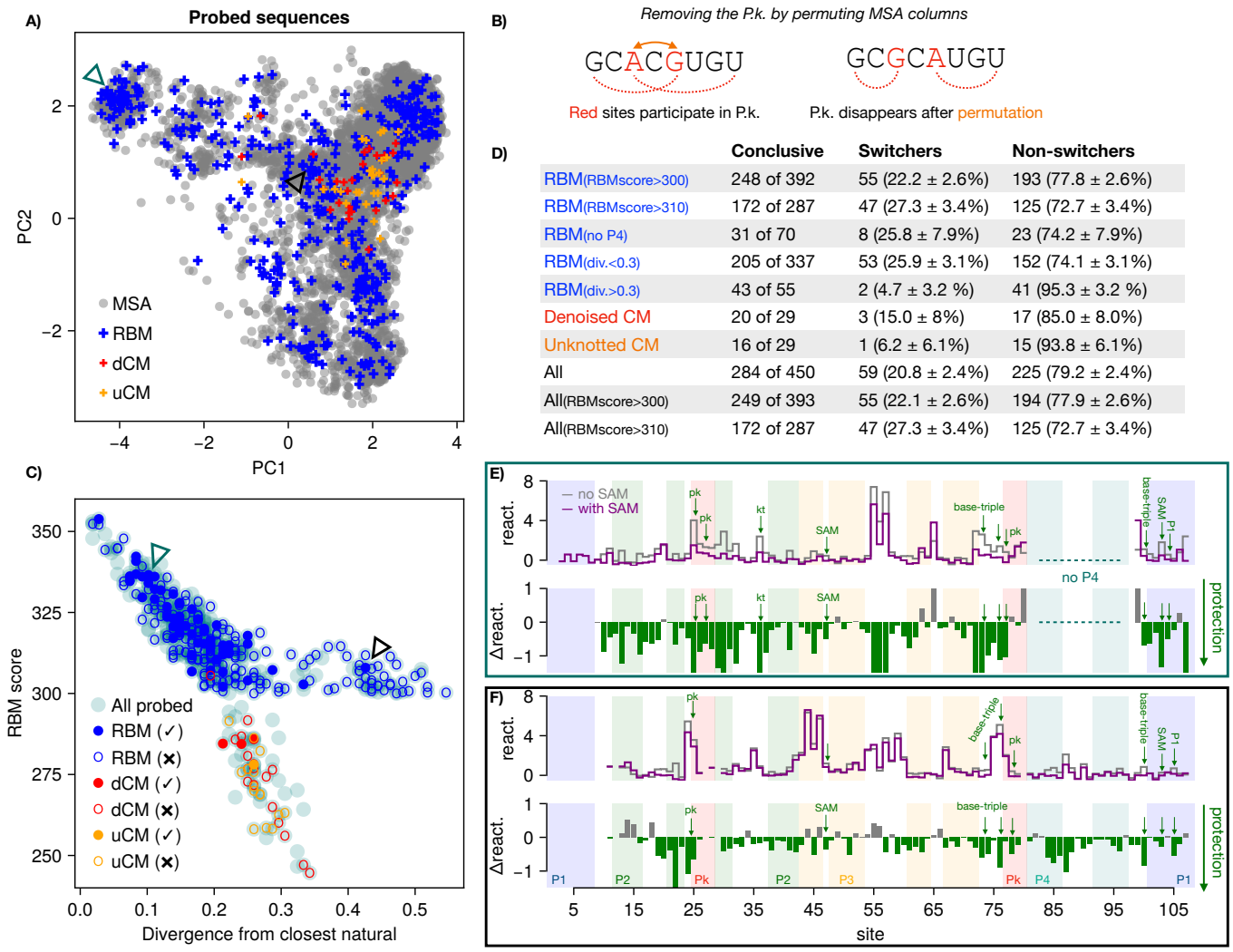
that stabilize Pk are also switchers. On the other hand, out of 112 identified switcher aptamers, in natural and artificial sequences, only 19 do not stabilize P1 significantly after binding SAM. These aptamers must exhibit significant compensatory stabilization of other structural Hallmark motifs from Table I. Taken together, these results are consistent with the known importance of the pseudoknot and P1 in the response of the aptamer in the context of the whole riboswitch.

D. Further explorations of RNA switch diversity through design

We then performed a second batch of design and experimental validation to further assess the limits of our generative models. We probed a total of 450 generated aptamers, whose sequences are projected onto the MSA PCs in Fig. 8A.

First, from the RBM model, we sampled sequences exhibiting higher distances from their closest natural counterpart, focusing on RBM scores > 300. Also, since we found a cluster of natural sequences lacking P4, we filtered a subset of RBM generated sequences having severely diminished P4 lengths. These are clearly seen in Fig. 8A, clustered at the top-left corner of the plot (recall the top PC1 represents P4 deletion). In addition to these, we also sampled more RBM sequences of high scores (> 300 and > 310), to obtain better statistics on the fractions of working aptamers.

Regarding CM, we considered two variations hoping to improve their generative properties. First, note that to be able to capture more distant sequences in the alignments, the Rfam CM models are regularized, resulting in a less constrained model (we have referred to this variant as *Rfam CM*). Consequently, we rebuilt a non-regularised CM trained on the same MSA, which we call *Denoised CM* (see Supplementary Fig. S10 and Methods). Secondly, since CM are unable to model pseudoknots, we devised a permutation of the MSA columns that undoes the pseudoknot, see Fig. 8B. We trained a new CM variant on the permuted MSA, that we call *Unknotted CM*, properly taking into account covariations in the pseudoknot; generate sequences with such model and finally



D)

	Conclusive	Switchers	Non-switchers
RBM(RBMscore>300)	248 of 392	55 (22.2 ± 2.6%)	193 (77.8 ± 2.6%)
RBM(RBMscore>310)	172 of 287	47 (27.3 ± 3.4%)	125 (72.7 ± 3.4%)
RBM(no P4)	31 of 70	8 (25.8 ± 7.9%)	23 (74.2 ± 7.9%)
RBM(div.<0.3)	205 of 337	53 (25.9 ± 3.1%)	152 (74.1 ± 3.1%)
RBM(div.>0.3)	43 of 55	2 (4.7 ± 3.2 %)	41 (95.3 ± 3.2 %)
Denoised CM	20 of 29	3 (15.0 ± 8%)	17 (85.0 ± 8.0%)
Unknotted CM	16 of 29	1 (6.2 ± 6.1%)	15 (93.8 ± 6.1%)
All	284 of 450	59 (20.8 ± 2.4%)	225 (79.2 ± 2.4%)
All(RBMscore>300)	249 of 393	55 (22.1 ± 2.6%)	194 (77.9 ± 2.6%)
All(RBMscore>310)	172 of 287	47 (27.3 ± 3.4%)	125 (72.7 ± 3.4%)

FIG. 8. Generation of additional sequences. **A)** Projection of sequences probed in second set of designed sequences along the PCs of the natural MSA, colored by origin: RBM, Denoised and Unknotted CM (dCM, uCM). The full natural MSA is shown in background (gray) for comparison. **B)** Diagram explaining the Unknotted CM, where the pseudoknot is undone by permuting a specific set of columns in the MSA. In this manner, a CM can model covariation along a pseudoknot. See Methods for details. **C)** Divergence from closest natural sequence in the MSA (fraction of sites that differ) vs. the RBM score, for all sequences probed in the second experiment. Colored circles correspond to aptamers that switch in response to SAM (fill color) or not (empty), with the color indicating the sequence origin: RBM (blue), dCM (red), and uCM (orange). Sequences for which our analysis is inconclusive are shown in light cyan. **D)** The table summarizes the numbers of switching sequences in each group. **E)** Reactivity profile of example switcher RBM generated sequence with no P4 (indicated by teal triangle in A,C). **F)** Reactivity profile of example switcher RBM generated sequence at large distance from natural sequences (indicated by black triangle in A,C).

permutate back the pseudoknot columns (See Methods).

Interestingly, both Denoised and Unknotted CM share some of the properties of Rfam CM noted previously. First, CM generated sequences from all variants have predominantly low RBM scores < 300, see Supplementary Fig. S11. Second, CM generated sequences exhibit restricted diversity, concentrating in a central region of the PCA plot, as in Fig. 4D. In particular, all CM are unable to generate sequences without the P4 helix. Sequences from the Unknotted CM have better complementarity and Turner energies favorable for base-pairing along the pseudoknot.

We then performed the same SHAPE-MaP experimental and analysis pipeline as in the previous sections. We summarize the results in Fig. 8C, which shows the RBM score of the probed aptamers against their Hamming distance to the closest natural sequence. Sequences that switch conformation in response to SAM are shown as filled disks.

Out of the 248 conclusive RBM sequences in the second

batch, 22% switch in response to SAM. The percentage of switchers among the sequences closer to the natural ones is higher and compatible to what obtained in replicate 1 considering the error bars, see Table 6C.

Moreover, 25% of the RBM aptamers having P4 lengths ≤ 1, respond to SAM, and an example reactivity profile is shown in Fig. 8E. We also find a few switching aptamers differing in between 30 to 50 sites from any natural sequence; an example reactivity profile is shown in Fig. 8F. The reactivity profile is compatible with the consensus secondary structure, with most reactivity peaks tending to occur in unpaired loops (except a portion of P3 that remains reactive), and an overall protection in response to SAM compatible with binding and stabilization of the aptamer.

In contrast, only 3 out of 20 conclusive Denoised CM samples switch in response to SAM (15%), and only 1 out of 16 from Unknotted CM ($\approx 6\%$). Thus the Denoised and Unknotted CM perform better than Rfam CM, but not as good as RBM.

In summary also for the second batch we find that a substantial fraction of RBM sequences switch conformation in response to SAM. Remarkably, RBM is able to generate novel functional aptamers that lack a P4 helix, and some of the more distant sequences (between 30 to 50 mutations out of 108 from any natural) remain functional. These results support the generalization ability of the RBM.

III. DISCUSSION

In this work, we focused on the design of molecular RNA switches, capable of changing conformation upon binding to a metabolite. Building such small RNAs has many potential applications in developing laboratory tools for gene function studies, metabolic engineering or drug design, as they can be used to regulate gene expression. The design of allosteric and regulatory RNA is also key to DNA-RNA computing, and to the investigation of possible scenarios for the origin of life [9, 33, 67].

State-of-the-art design methods for RNA are based on computational frameworks to fold sequences in a given secondary structure from the knowledge of thermodynamics parameters for the pairing energies [81], possibly including tertiary elements such as pseudoknots [87]. Such methods have been used to obtain sequences with bistable secondary structures [25] and extended to take into account both positive and negative design elements [58, 87], as well as to participative rational design [39]. Our design method, based on the unsupervised generative architecture of Restricted Boltzmann Machines, differs in two key ingredients: i) it exploits the sequences (of SAM-I riboswitch aptamers) sampled through evolution and collected in data bases, building upon the frameworks introduced in homology and covariation detection [15, 48, 53, 58, 86]; ii) it encompasses, through learning of a unique parametric model, the arrangements of nucleotide motifs allowing natural sequences to acquire adequate secondary and tertiary structures and to undergo an allosteric response to metabolite binding.

We have verified that the RBM model learned from sequence data encode nucleotide-nucleotide contacts in the secondary structure and in the pseudoknot, performing at the same level as pairwise Potts/DCA models previously introduced to this aim [15, 86]. However, contrary to those pairwise interaction based models, RBM are capable of extracting extended nucleotide motifs, *e.g.* overlapping one or more structural elements. In addition, we stress that a major advantage of the shallowness of the RBM architecture is that the repertoire of these motifs can be readily accessed and interpreted through inspection of its weights (Fig. 3B&C).

To assess the sequences designed by our computational models, as well as the natural sequences belonging to the SAM-I riboswitch aptamer family, we have carried out high-throughput SHAPE-MaP [69] screening. We have developed and implemented a statistical pipeline to analyze the measured SHAPE reactivities, based on a likelihood ratio between reactivity distributions of paired/unpaired nucleotides (that we called protection score) [22]. Our analysis takes advantage of the closely related statistics of the ensemble of tested sequences and their shared consensus secondary structure. Importantly, it does not rely on a biophysical implementation of the Turner model [40]. Tertiary contacts such as pseudoknots, which are essential to model complex conformational changes such as those occurring in riboswitches but pose difficulties to other tools [17, 64], are here fully accounted for. Last of all, our pipeline for SHAPE-MaP screening is fully automatic and does not require manual annotation, which is time consuming for high-throughput screening.

The outcome of SHAPE-MaP shows that RBM models are able to design artificial SAM-I-riboswitch-like aptamers that successfully switch between open and closed conformations upon SAM binding. Out of the high-score generated sequences for which our conservative statistical analysis could reach a clear conclusion, 35% could be classified as switchers in the first replicate. This fraction is significant, and shows that RBM are effective as generative models of complex RNAs. It is however lower than the one (70%) of natural sequences deemed as switchers according to the same criterion. We emphasize that the fraction quoted above varies with the constraints considered during the generation process. For instance, up to 50% of RBM-generated sequences were recognized as switchers when the average distance to the closest natural sequences in the MSA (defined as the frequency of mutations across the 108 nucleotides) is of 20%. Pushing generation to the limits as in the second experiment made the global fraction drop down to 22%, but allowed us to generate functional switchers with as many as 46% of mutations with respect to the closest known natural aptamers. Moreover, RBM can design switcher aptamers lacking the P4 helix (as in some natural variants),

whereas CM are unable to generate such sequences.

The success of our design approach crucially relies on the capability of RBM to capture nucleotide motifs responsible for tertiary structural elements. This statement is supported by the fact that CM, while capturing the local conservation and secondary structure of the Riboswitch family, has significantly lower generative performance ($\simeq 11\%$, Denoised & Unknotted). In addition, RBM generate flexible structural elements, with intermediate pairing energy values, permitting them to open and close depending on the metabolite presence. From this point of view, while RBM have already been used to generate functional proteins [43] or DNA aptamers [19], this is the first time they are shown to be able to design allosteric biomolecules.

Besides the switcher/non-switcher classification, a pattern of phenotypes is observed in the generated sequences through manual inspection of the reactivity profiles and of their changes with SAM presence. Among the natural sequences that fail to qualify as switchers with our automatic statistical pipeline, many are manually recognised as responsive to SAM (Fig. 6H). This response can manifest itself as a change in the reactivities of the sites related to the SAM binding pocket, or involved in P1, in the pseudoknot, or in any of the three base triples. Similar patterns are encountered in RBM-generated sequences, see Supplementary Section J. This result suggests that some natural sequences in the SAM I riboswitch RFAM family are indeed non-switchers under our experimental conditions but do bind SAM. Consequently, the RBM trained on those data can generate synthetic RNA sequences with a similar behaviour. Notice that the distinction between responsive to SAM and capable of undergoing conformational change we observe here agrees with recent directed evolution experiments. It was reported that evolving RNA for ligand binding alone often failed to produce functional regulatory RNAs [32, 84], highlighting the importance of the structural switch. More recently, Capture-SELEX, in which conformational change triggered by the ligand is selected for, has been proposed for this purpose [3].

Since this paper was posted on the archive, two works have developed generative models of structured RNA: [7] proposes a parsimonious DCA-like model, which promotes sparsity of model weights and validated experimentally generation of a tRNA family; [74] introduced a combination of Variational AutoEncoders with CM and showed that their model was generative over various rybozyme families. Our work differs in that it presents the first example of design of RNA molecules exhibiting structural switching upon metabolite binding.

We plan to extend our approach to the modeling of complete SAM riboswitches by including the expression platform, and to investigate more deeply the mechanisms for functionality switches in different subfamilies of the SAM-riboswitches family. Finally the RBM model could also be used to design other RNA based on their respective MSA, including longer and more complex RNA such as ribosomal and messenger RNA.

IV. MATERIALS AND METHODS

A. Multiple sequence alignment of SAM-I riboswitches

The RF00162 family from the Rfam database [36] groups sequence homologs of the aptamer domain of the SAM-I riboswitch. We downloaded a manually curated seed alignment from Rfam (version 14.7), containing 457 aptamer sequences supported by literature evidence. These seed sequences are aligned to a consensus secondary structure (shown in Fig. 1B) that has been informed by the holo-form of SAM-I riboswitch crystal structures [41, 49]. After removing extended stems and variable loops, labeled as insertions in the alignment, we obtain 108 matched positions (including gaps that mark deletions) spanning four helices that interleave around a central four-way junction. We trained a covariance model (CM) [21] on this seed alignment using Infernal [52] with default settings. Following standard protocols [35], we then acquired 6161 additional sequences from Rfam, collected from genome databases search fetching for significant matches to the CM model. We constructed a multiple sequence alignment (MSA) containing these sequences, that we will refer to below as the full MSA, to distinguish from the seed MSA that consists only of the 457 manually curated seed sequences. The sequence conservation logo of the full MSA is shown in Fig. 1A.

B. Infernal pipeline

Infernal [52] is a set of computational tools to facilitate modelling RNA sequence families under a profile stochastic context-free grammar (pSCFG) formalism, also known as covariance models (CM) [21]. A CM is capable of modelling the conservation profile of important sites along the sequence, as well as correlations between distant sites required by the complementarity of base-pairs in a given secondary structure. Infernal is routinely used in the maintenance of alignments in the Rfam database [35, 36]. We employed Infernal to construct the RF00162 full MSA, that we use to train the RBM.

By restricting to covariations in the secondary-structure, CM enable implementation of efficient dynamical programming algorithms to train, sample, and scan large genomes for sequences that score highly under the CM [21]. However, these assumptions also imply that the CM is unable to include additional constraints in the probabilistic sequence model, such as pseudoknots and other tertiary contacts in the 3-dimensional fold of the RNA molecule.

1. Rfam CM.

The Rfam database associates a CM model to each family, trained on the seed alignment, that is used to efficiently scan large genomes for significant sequence matches to the family (the hits). The raw CM model downloaded from Rfam is significantly regularized so that it is more effective in fetching far homologs of a family in deep genome searches [51]. We will refer to this CM model as the *Rfam CM*.

2. Denoised CM.

Since the *Rfam CM* is strongly regularized, in this work, we also trained a CM model variant on the full MSA, with no regularization, that we call *Denoised CM*. This model reproduces more closely some statistics of the full MSA (conservation and covariances associated with the secondary structure), possibly at the expense of recognizing a more restricted set of homologs.

3. Unknotted CM.

As discussed previously, a CM model is unable to model pseudoknots and other tertiary contacts in the 3-dimensional structure of a RNA molecule. Based on our knowledge of the consensus secondary structure of the SAM-I riboswitch aptamer (Fig. 1A), we devised a third CM model able to account for sequence covariation in pseudoknot sites constructed as follows. Columns 77–80 of the MSA, corresponding to the sites on the 3'-end part of the pseudoknot, were moved and inserted after site 28, right next to the the sites at the 5'-end of the pseudoknot. In this way, the pseudoknot is "unknotted", and is now representable in the CM model as part of a pseudo-secondary structure corresponding to the permuted MSA. Accordingly, we proceeded to train a new CM model on the rearranged full MSA. We call the resulting model the *Unknotted CM* in what follows.

4. Sampling the CM.

To better understand the limitations of CM models and the advantages of RBM, we sampled 10000 sequences from each of the three CM described above. For the Unknotted CM, the rearranged columns are permuted back to their original positions after sampling. We used Infernal's `cmemit` program with default parameters, and without insertions. Infernal computes a score of sequences aligned to the CM, related to the likelihood of the CM to emit a given sequence (also called *bit-scores*). We computed this score using `cmalign`, with `-g` (global) option to avoid local approximations [51].

C. Restricted Boltzmann machines

1. Definitions.

Restricted Boltzmann Machines (RBM) [31] are bipartite graphical models over N visible variables $\mathbf{v} = \{v_1, v_2, ..., v_N\}$ and M hidden (or latent) variables $\mathbf{h} = \{h_1, h_2, ..., h_M\}$, see Fig. 2A. In sequence data modelling, N corresponds to the sequence length ($N = 108$ for the SAM-I riboswitch), and the values of v_i encode for the nucleotide present at position i of the sequence. In the case of RNA, v_i can take one of $q = 5$ possible values, corresponding to the nucleotides A, C, G, U, and the gap symbol ($-$) of the alignment. The hidden variables h_μ can be real-valued, and their number M will determine the capacity of the machine to model complex correlations in the data, as we will see below. The two layers are connected through the interaction weights $w_{i\mu}$. An RBM defines a joint probability distribution over $\mathbf{v}$ and $\mathbf{h}$ through

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}, \tag{1}$$

where Z is a normalization factor, known as the partition function, and the energy $E(\mathbf{v}, \mathbf{h})$ is given by

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^{N} \mathcal{V}_i(v_i) + \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) - \sum_{i=1}^{N}\sum_{\mu=1}^{M} w_{i\mu}(v_i) h_\mu \tag{2}$$

The functions $\mathcal{V}_i(v_i)$, $\mathcal{U}_\mu(h_\mu)$ are potentials biasing the distributions of single units. The visible units v_i can take a finite number of possible values, and therefore the quantities $\mathcal{V}_i(v_i)$, also called 'fields', can be stored as a $q \times N$ matrix. Similarly, the weights $w_{i\mu}(v_i)$ can be stored as a $q \times N \times M$ three-dimensional tensor. The hidden variables, on the other hand, are continuous, and we chose to parameterize their potentials using the double Rectified Linear Units (dReLU) form, proposed previously in [80],

$$\mathcal{U}_\mu(h_\mu) = \begin{cases} \gamma_\mu^+ h_\mu^2/2 - \theta_\mu^+ h_\mu & h_\mu \geq 0 \\ \gamma_\mu^- h_\mu^2/2 - \theta_\mu^- h_\mu & h_\mu \leq 0 \end{cases} \tag{3}$$

with real parameters $\gamma_\mu^\pm, \theta_\mu^\pm$, satisfying $\gamma_\mu^\pm > 0$. The dReLU is an attractive choice because it is expressive enough to cover several interesting settings. When $\gamma_\mu^+ = \gamma_\mu^-$ and $\theta_\mu^+ = \theta_\mu^-$, Eq. (3) becomes a quadratic (*i.e.*, Gaussian) potential, and is closely related to Direct-Coupling Analysis models popular in protein sequence modelling [11, 15, 50, 63, 68, 85]. However, the Gaussian choice is unable to parameterize more than two-body interactions, which can be a limitation in RNA structure where some interactions are known to involve more than two sites (*e.g.* stacking interactions [10, 88]), as well as functional interactions that can span complex structural and sequence motifs. dReLU can also adopt a bimodal form when $\theta_\mu^+ > 0 > \theta_\mu^-$, and more generally are able

to capture extensive sequence motifs, beyond two-body interactions.

The likelihood of visible configurations under the RBM can be obtained by marginalizing over the states of the hidden units:

$$P(\mathbf{v}) = \frac{1}{Z}\int e^{-E(\mathbf{v},\mathbf{h})}d\mathbf{h} = \frac{1}{Z}e^{-E_{\text{eff}}(\mathbf{v})} \tag{4}$$

where $-E_{\text{eff}}(\mathbf{v})$ is the resulting RBM score, as a function of visible configurations $\mathbf{v}$ only, that incorporates effective interactions arising from the marginalized latent variables (see Fig. 2C):

$$E_{\text{eff}}(\mathbf{v}) = \sum_{i=1}^{N}\mathcal{V}_i(v_i) - \sum_{\mu=1}^{M}\ln\int e^{\sum_i w_{i\mu}(v_i)h_\mu - \mathcal{U}_\mu(h_\mu)}dh_\mu \tag{5}$$

Although evaluating $P(\mathbf{v})$ is computationally difficult (because the partition function Z is intractable), Eq. (5) shows that the score $-E_{\text{eff}}(\mathbf{v})$ can be computed efficiently.

2. *Training the RBM.*

The likelihood assigned by the RBM to a sequence depends on all the parameters of the model: the $q \times N \times M$ weights tensor $w_{i\mu}(v_i)$, the $4M$ hidden unit dReLU parameters $\gamma_\mu^\pm, \theta_\mu^\pm$, and the $q \times N$ visible unit fields $\mathcal{V}_i(v_i)$. Given a set of aligned data sequences, these parameters are learned, in an unsupervised way, by maximizing the average log-likelihood of the data

$$\mathcal{L} = \frac{1}{B_{\text{MSA}}}\sum_{\mathbf{v}\in\text{MSA}}\ln P(\mathbf{v})\,, \tag{6}$$

(where the sum is taken over the B_{MSA} sequences in the MSA) plus a regularization term,

$$\mathcal{R} = -\frac{\lambda_{\text{reg}}}{2}\sum_{\mu=1}^{M}\left(\sum_{i=1}^{N}\sum_{v=1}^{q}|w_{i\mu}(v)|\right)^2, \tag{7}$$

(where λ_{reg} is a non-negative parameter). This form of the regularization, combining L_2 and L_1 norms, has been proposed by [80] to favor sparse weights with balanced norms across all hidden units. Regularization helps avoid over-fitting and promotes a smoother training and sampling of the model [31].

To train the model, we perform a variation of gradient ascent over $\mathcal{L} + \mathcal{R}$. Schematically, if ω_t denotes a parameter of the RBM (weights $w_{i\mu}$ or a parameter of the potentials $\mathcal{V}_i, \mathcal{U}_\mu$), then after t iterations of the optimization, the parameters are updated during learning as follows:

$$\omega_{t+1} = \omega_t + \eta\frac{\partial}{\partial\omega}(\mathcal{L} + \mathcal{R})\,, \tag{8}$$

where η is a small positive learning rate. In practice, the optimization is accelerated by an adaptive momentum term [38] and a centering trick [46], following the implementation of RBM in [23]. See Supplementary note A for further details.

Computing the gradient of $\mathcal{L}$ requires to estimate the moments of visible and/or hidden variables with respect to the model distribution [31]. We employ the Persistent Contrastive Divergence (PCD) algorithm [77], where a number of Markov chains sampled from the model are updated in each parameter update. We have that:

$$\frac{\partial\mathcal{L}}{\partial\omega} = \underbrace{\left\langle\frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial\omega}\right\rangle_{\text{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle\frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial\omega}\right\rangle_{\text{RBM}}}_{\text{negative gradient}}. \tag{9}$$

The first term is an empirical average performed over the data, as in Eq. (6), while the second term is averaged over sequences sampled from the RBM. We represent these terms as arrows in Fig. 2B. The first term (blue), tends to drive the parameters ω of the RBM such that the score $-E_{\text{eff}}(\mathbf{v})$ of sequences $\mathbf{v}$ in the data is increased. This results in the model assigning higher probabilities to regions of sequence space densely populated by data. To do this, the hidden units of the RBM must extract features shared by the data sequences and thus likely to be important for their biological function. Conservation of probability implies that regions of sequence space not populated by data sequences must be penalized. This is taken care of by the second term in Eq. (9) (in red), which tends to increase the average energy of sequences uniformly sampled from the RBM. The net effect of the two terms (also called positive and negative phase terms in earlier papers [31, 65]), is that the RBM places most probability mass in regions densely populated by data and low probability elsewhere. However, the finite parameterization and discovered features usually extrapolate also to novel regions in sequence space, not covered by the data, where the model assigns high probability, as illustrated in green in Fig. 2B. The trained RBM model automatically extracts features and constraints from the data, which are then imposed in the generated sequences, in a manner akin to the features used for *positive* and *negative design*, in rational design approaches.

3. *Sampling the RBM.*

Having trained the model, sampling can be performed through a Monte-Carlo procedure known as Gibbs sampling. It exploits the two-layer RBM architecture, by noting that the conditional distributions of one layer given

the configuration of the other layer, factorize:

$$P(\mathbf{v}|\mathbf{h}) \propto \prod_{i=1}^{N} \exp\left(-\mathcal{V}_i(v_i) + \sum_{\mu=1}^{M} w_{i\mu}(v_i)h_\mu\right)$$
$$P(\mathbf{h}|\mathbf{v}) \propto \prod_{\mu=1}^{M} \exp\left(-\mathcal{U}_\mu(h_\mu) + \sum_{i=1}^{N} w_{i\mu}(v_i)h_\mu\right) \tag{10}$$

These conditional distributions are therefore easy to sample. The Gibbs sampling algorithm consists of iteratively sampling one layer conditioned on the other layer, and vice-versa, for a number of steps, and collecting the configuration at the final iteration. If a large enough number of steps are taken, the resulting sample is guaranteed to be a good approximation of an equilibrium sample of the RBM. Equilibration can be assessed by inspecting convergence of quantities such as the average energies of the samples. For the RBM we trained in this work, we found that ~ 1000 Gibbs sampling steps were more than sufficient to reach equilibirum.

4. *Estimating the partition function.*

Evaluating the likelihood $P(\mathbf{v})$ of a given test sequence $\mathbf{v}$ requires computing the partition function Z in Eq. (1), which involves an untractable summation over all the possible q^N sequences. An approximate Monte-Carlo scheme known as annealed importance sampling allows us to obtain estimates of Z [54]. Our implementation follows [23].

5. *Epistatic scores.*

The riboswitch aptamer structural fold imposes contacts between distant sites along the RNA sequence, which are reflected in covariations between nucleotides in the corresponding columns of the MSA. To assess how well the RBM captures sequence features connected to these structural constraints, we compute the following epistatic score [80]:

$$\mathcal{J}_{ij} = \sum_{a,b} \left\langle \frac{1}{25} \sum_{a',b'} \ln\left[\frac{P(\mathbf{v}_{ij}^{a,b})P(\mathbf{v}_{ij}^{a',b'})}{P(\mathbf{v}_{ij}^{a',b})P(\mathbf{v}_{ij}^{a,b'})}\right]\right\rangle_{\text{MSA}}^{2} \tag{11}$$

for pairs of sites i, j along the sequence. Here, $\mathbf{v}_{ij}^{a,b}$ denotes a sequence $\mathbf{v}$ from the MSA, which has suffered a double mutation: site i was modified to symbol a, and site j was modified to symbol b. $P(\mathbf{v}_{ij}^{a,b})$ denotes the likelihood, Eq. (4), of this modified sequence, and the average $\langle \ldots \rangle_{\text{MSA}}$ is taken over all sequences of the MSA. Note that we average over all possible pair of mutations a', b' at sites i, j, dividing by $25 = 5^2$, the number of possible letters (4 nucleotides and a gap symbol) at these two positions. This score, introduced by [80], is closely related to the Frobenius norm of interactions used in Direct-Coupling analysis for contact prediction in proteins [11], and measures how the epistatic effect of a pair of mutations is enhanced in comparison to the effects of the single mutations by themselves. In addition, we apply the average-product correction (APC) to the matrix $\mathcal{J}_{ij}$, which has been argued to decrease the impact of phylogenetic biases in contact prediction [11, 20]. The APC corrected contact matrix, $\tilde{\mathcal{J}}_{ij}$, is defined by:

$$\tilde{\mathcal{J}}_{ij} = \mathcal{J}_{ij} - \frac{\sum_{kl} \mathcal{J}_{kj}\mathcal{J}_{il}}{\sum_{kl} \mathcal{J}_{kl}} \tag{12}$$

6. *Further details.*

We trained an RBM with 108 Potts visible units corresponding to the aligned sequence sites, and 100 hidden dReLU units, with a regularization weight of $\lambda_{\text{reg}} = 0.01$. Our implementation of the RBM training, sampling and partition function calculations, follows closely that of [23]. Additional details can be found in the Supplementary Note A. In addition, we have conducted detailed cross-validation experiments supporting our architectural choices, see Supplementary Fig. S1.

D. Biophysical energy calculations

We computed biophysical pairing energy predictions for the formation of P1 and the pseudoknot of various sequences using the Turner energy model, as implemented in the ViennaRNA package [40], with the RNAeval program. For the P1 helix, we computed the energy difference of each sequence in the consensus secondary structure, where P1 is paired (Fig. 1B), and in a conformation where P1 is unpaired (Fig. 1A).

To estimate the pairing energy associated to pseudoknot formation, we used RNAeval on a virtual secondary structure where only the pseudoknot sites are base-paired, and all other sites are unpaired. We then considered the interior loop contributions to the resulting folding energy.

Note that in both cases, intrinsic limitations of the ViennaRNA implementation imply that we cannot model the pseudoknot together with other structural elements (and other tertiary contacts) simultaneously.

E. Selection of sequences for first batch

We probed a total of 306 sequences breaking down as follows.

RBM sequences. We generated sequences from the RBM by Gibbs sampling. Equilibration was assessed by monitoring the average score of the sample. We found that 5000 steps were more than sufficient. We then sorted these sequences by their value of $-E_{\text{eff}}$, and selected 70

sequences at random, uniformly spanning the range of scores observed in the sample. The table of sequences and their associated RBM scores is reported in the Supplementary Materials.

We hypothesized that functionality of RBM sequences should strongly correlate with their RBM score, $-E_{\text{eff}}(\mathbf{v})$. We therefore define a group having *high*-RBM scores, with $-E_{\text{eff}}(\mathbf{v}) > 300$ and consisting of 53 sequences. We also define a group of *low*-RBM score sequences, having $-E_{\text{eff}}(\mathbf{v}) < 300$, and containing 31 sequences. Setting score thresholds is similar to sampling the model at a temperature different from 1, also found to be helpful in DCA models used for design [63]. The chosen threshold value $-E_{\text{eff}} = 300$ approximately separates the bulk of natural and RBM sampled sequences, from the CM samples (Fig. 4A). We also emphasize that the RBM automatically tends to sample predominantly sequences with $-E_{\text{eff}}(\mathbf{v}) > 300$.

Infernal sequences. We also sampled sequences from the Rfam CM covariance model of the RF00162 family, downloaded from Rfam. We used the Infernal cmemit program (see Methods) to sample a large batch of sequences. Then, we selected 30 sequences uniformly spanning the range of Infernal bit-scores observed during sampling. The covariance model can model correlations along the sequence comprised in the consensus secondary structure, but fails to account for other correlations arising from tertiary contacts or pseudoknots. Therefore, these Infernal generated sequences serve as a baseline.

Natural aptamers. We selected 151 sequences members of the seed MSA and 55 sequences members of the full MSA, as described in SectionIV A. The selected natural sequences are diverse, spanning various taxonomic classes (see Fig. 4D). A listing of probed sequences can be found in the Supplementary Material.

F. Selection of sequences for second batch

In the second experiment, we generated a total of 450 sequences to be probed, of different origins. We considered:

- 58 CM sequences, with 29 from Unknotted CM and 29 from Denoised CM (see Section IV B for definitions of these CM variants).
- 392 sequences sampled from the RBM, filtered to have RBM scores > 300. In particular, 49 of them were selected because they had no P4 helix, while 100 of them were selected because they had larger Hamming distances from any natural sequences.

The full list of designed sequences is provided as part of the Supplementary Code listing, see Section IV L.

G. SHAPE-MaP experiments

1. RNA preparation.

DNA oligonucleotides representing the 206 SAM-I natural sequences, and the two batches (100 and 450) of artificial sequences, preceded by the T7 promoter (5'CGGCGAATCTAATACGACTCACTATAGG3') and followed by a tag sequence representing a 10 nucleotide barcode unique for each aptamer and a primer binding site, were purchased as an oligonucleotide pool (Twist bioscience®). The Tag sequence was designed to avoid interference with the aptamer secondary structure using RNAFold [40]. The oligo pool was PCR amplified using the T7 promoter as forward primer and and five different reverse primers. RNA was transcribed and prepared as previously described [16], and was checked for the absence of aberrant products on a 1% agarose gel.

2. SHAPE probing.

SHAPE chemical probing was performed as described previously [69]. Briefly, 10 pmol of RNA were diluted in 12 µL of water and denaturated for 3 min at 85°C. Then, 6µL of 3X pre-warmed folding buffer with or without magnesium (0.3M HEPES pH 7.5, 0.3M KCl, 15mM MgCl2) were added and the solution was allowed to cool down to room temperature. Samples were then incubated at 30°C for 5 min. S-adenosyl-methionine (SAM) was added at final concentration of 0, 0.1 or 1mM and samples were incubated 15 min at 30°C. 9uL (corresponding to 5 pmoles) were aliquoted and 2uL of 50 mM 1M7 (1-Methyl-7-nitoisatoic anhydre) or DMSO (Mock reaction) was added and allowed to react for 6 min at 30°C. RNAs were then reverse transcribed with the Superscript III reverse transcriptase (Invitrogen®) and NGS libraries were prepared using NEBNext Ultra II DNA Library Prep Kit (New England Biolabs®). Final products were sequenced by using the illumina technology (NextSeq 500/500 Mid 2x150 flowcell). Sequencing data were analyzed and reactivity maps were derived using ShapeMapper2 [6]. In the end, the 306 selected sequences were probed in the following conditions:

- 30°C, without Mg^{2+} and without SAM.
- 30°C, with magnesium (Mg^{2+}).
- 30°C, with magnesium and 2 concentrations (0.1 and 1mM) of SAM.

Each probing reaction was repeated in triplicate. The 2 SAM concentrations were analyzed together, since we found no significant effect of varying the SAM concentration in the reactivity responses of the aptamers (see Supplementary Fig. S15) and in this manner we had more statistics.

3. Manual inspection of reactivity profiles.

IPANEMAP [64] was used to generate RNA secondary structure models for each sequences. For manual inspection, we considered the reactivity of the nucleotides known to be directly involved in SAM binding (U7, G11, A46, U69, G70, U103) and of those known to be protected from shape reactivity in the closed stated, *i.e.*, nucleotides in P1 (1-8; 101-108), in the pseudoknot (25-28; 77-80), those involved in the three base triple interactions (24, 73, 74, 76, 100). Nucleotide numbering is after the consensus sequence shown in Fig. 1. An aptamer was considered to bind SAM if at least three of these elements are noticeably less reactive upon SAM addition, and if none of the binding determinant remain highly reactive. Note that P1 and the Pk are each considered as one element, and that some of the elements may be unreactive even in absence of SAM.

H. Statistical analysis of SHAPE-MaP reactivities

1. Reactivity definition.

SHAPE-MaP experiments result in measurements of sequencing error rates at each site of the RNA sequence, that correlate to the locations where the SHAPE probe has reacted with the RNA. For each site $i = 1, \ldots, N$ of a sequence n, the reactivity is defined by [69]:

$$r_{in} = \frac{m_{in} - u_{in}}{d_{in}} \tag{13}$$

where m_{in} is the mutation rate in presence of the reagent, u_{in} is the mutation rate in its absence accounting for mutational background of the experiment, and d_{in} is the mutation rate in a denaturating condition where the RNA is expected to be unfolded, intended to cancel sequence-dependent biases. Working with r_{in} is usually better since this form should cancel site-dependent biases in the raw SHAPE mutation rates, m_{in}. The basis of the SHAPE-MaP procedure relies on differences in the distribution of reactivities in base-paired and unpaired sites [69]. We have confirmed such differences are observed in our data in Fig. 5 (and also Supplementary Fig. S2).

2. Statistical analysis.

The finite number of sequencing reads collected at a site implies a statistical error in the reactivity computed by Eq. (13). Therefore, we cannot directly access the true reactivity r_{in} at a site, but rather an experimental measurement $\tilde{r}_{in}$ that fluctuates according to the number of reads taken at the site. To model this uncertainty, we make the simplifying assumption that the ideal reactivity of a site, r_{in}, depends only on whether the site is base-paired or not. Under this assumption, we can write:

$$\frac{P_{in}(\tilde{r}_{in}|\text{bp})}{P_{in}(\tilde{r}_{in}|\text{np})} = \frac{\int P(r|\text{bp}) P_{in}(\tilde{r}_{in}|r) dr}{\int P(r|\text{np}) P_{in}(\tilde{r}_{in}|r) dr} \tag{14}$$

where:

- $P_{in}(\tilde{r}_{in}|\text{bp})$ is the probability of measuring a reactivity $\tilde{r}_{in}$ at site i of sequence n, given that the site is base-paired and conditioned on the finite number of reads taken at this position.
- $P_{in}(\tilde{r}_{in}|r)$ is the probability of measuring reactivity $\tilde{r}_{in}$ at site i of sequence n, on account of fluctuations due to a finite number of reads, conditioned on this site having a real reactivity of r.
- $P(r|\text{bp})$ is the probability distribution of reactivities of base-paired sites, at infinite read-depth, assumed to be homogeneous across sites.
- $P_{in}(\tilde{r}_{in}|\text{np})$ and $P(r|\text{np})$ are defined in a similar manner for non-paired sites.

We approximate the distributions $P(r|\text{bp})$ and $P(r|\text{np})$ by kernel density estimators fit on the corresponding empirical histograms (shown in Fig. 5A for the first experiment). The kernel function used corresponds to a standard normal, with a bandwidth set according to the Silverman rule [70]. To better estimate the histograms, we use the experimental conditions with SAM, where the secondary structure of the aptamer is expected to be more stable. We also find that these histograms can depend on the particular experiment, and therefore we fitted $P(r|\text{bp}), P(r|\text{np})$ for each replicate.

Applying Bayes theorem [42] in Eq. (14), we can write:

$$\frac{P_{in}(\tilde{r}_{in}|\text{bp})}{P_{in}(\tilde{r}_{in}|\text{np})} = \frac{\int (P(r|\text{bp})/P(r)) P_{in}(r|\tilde{r}_{in}) dr}{\int (P(r|\text{np})/P(r)) P_{in}(r|\tilde{r}_{in}) dr} \tag{15}$$

where $P(r)$ is the histogram of real reactivities, regardless of whether a site is paired or not. The posterior $P_{in}(r|\tilde{r}_{in})$ quantifies the uncertainty of the real reactivity r at site i of sequence n, conditioned on our information of the measurement taken at this site. This uncertainty arises from the finite sequencing reads available, which induce an experimental error in our estimate of the quantities m, u, d appearing in Eq. (13). Since the mutation count at a site can be modeled by a Poisson distribution [69], the posterior distributions of the mutation rates m, u, d are Gamma distribution, assuming a convenient choice of conjugate prior [42]. Then, we can produce a Monte-Carlo estimate of $P_{in}(r|\tilde{r}_{in})$ by sampling the posterior Gamma distributions of m, u, d, and computing the reactivity through Eq. (13). If the sampled reactivities fall predominantly far in the tails of the histograms $P(r|\text{bp})$ or $P(r|\text{np})$, respectively, the reactivity measurement is discarded as an outlier. In practice, we find that 1000 samples for each site are more than sufficient. These samples can then be used to approximate the numerator and denominator of the right-hand

side of Eq. (15). In this way, we produce estimates of the ratios $P_{in}(\tilde{r}_{in}|\text{bp})/P_{in}(\tilde{r}_{in}|\text{np})$, quantifying the odds that a site is paired. Figure S17B shows a scatter plot of reactivities in our dataset, with the standard-error estimated by the standard SHAPE-Mapper pipeline [69] (which does a first-order error propagation through the Poisson count statistics), with each point colored according to the value of the log-odds-ratio Eq. (15). The dashed lines are approximate contours separating points which are over two-times more likely to be paired (blue) or unpaired (red). The fact that these contours are not straight vertical lines indicates that by using Eq. (15), we are considering both the reactivity value and its uncertainty in assessing the plausibility that a site is paired or not. A similar approach has been proposed by [22].

3. *SHAPE protection score of a structural motif.*

We can exploit the likelihood ratios $P_{in}(\tilde{r}_{in}|\text{bp})/P_{in}(\tilde{r}_{in}|\text{np})$ computed above to estimate the probability of the presence of a structural motif in a sequence. To be precise, by a motif of length $2L$, we intend a set of base-paired sites, $\mathcal{M} = \{i_1, j_1, \ldots, i_L, j_L\}$. For example, the P1 helix motif involves the following base-paired sites: $\{1, 108, 2, 107, \ldots, 8, 101\}$. We can then estimate the protection score that motif $\mathcal{M}$ is present in sequence n, from the reactivity data, as follows:

$$\mathcal{S}_n(\mathcal{M}) = \sum_{i \in \mathcal{M}} \ln \left(\frac{P_{in}(\tilde{r}_{in}|\text{bp})}{P_{in}(\tilde{r}_{in}|\text{np})} \right) \tag{16}$$

which we take as the definition of $\mathcal{S}_n(\mathcal{M})$ that we will use in what follows. One advantage is that this approach allows us to combine multiple reactivity measurements in a more robust probabilistic measure, achieving more statistical power than if we regarded individual site reactivities by themselves. This way, we can assess in a probabilistic manner, the presence of structural elements in our probed sequences n.

4. *Limitations and robustness.*

We acknowledge some limitations of the pipeline described in this section. First, the definition of the histograms of base-paired and unpaired reactivities (red and blue in Fig. S17A) rely on the consensus secondary structure of the SAM aptamer. However, it is well known that the riboswitch is flexible and the set of base-paired residues will depend on condition and aptamer.

To further validate the robustness of this pipeline, we have also repeated the analysis, varying the following settings:

- To assess the impact of wrongly annotated sequences, we replaced the histograms 5A, originally computed on all natural sequences, by analogous histograms computed on seed natural sequences only in the alignment. The resulting histograms are very close to 5A, see Supplementary Fig. S2B.
- One can argue that the histograms in Fig. 5A might be convoluted by noise, and further improvements could be obtained by attempting to deconvolve this noise. To assess the impact of the noise, we can go in the opposite direction, of further convolving $P(r|\text{bp})$ and $P(r|\text{np})$ with experimental noise and inspecting the effect on the resulting histograms. More precisely, we can resample the reactivities Eq. (13) from the site distributions $P_{in}(r|\tilde{r}_{in})$ and recompute the histograms using these resampled reactivities in place of the original ones. The resulting histograms suffer minor variations in comparison to 5A, see Supplementary Fig. S3).
- Sites labeled as base-paired / unpaired in the consensus secondary structure can be regarded as being so in only some conditions, and for some aptamers. Two examples are the P1 helix and the pseudoknot, which are generally expected to undergo rearrangements related to SAM binding. In the Supplementary Materials, we have performed the following experiments: i) exclude P1 from both histograms; and ii) include the pseudoknot in the unpaired histogram. See Supplementary Fig. S2C,D. In both cases, the histograms suffer minor variations.

As a consequence, despite these limitations, our conclusions remained unaffected in all the cases we tested.

5. *Error bars for the rates of switching aptamers*

Given $N_{conc.}$ conclusive probed sequences, N_{switch} of which are found to perform the structural switch, we estimated the success rate by $p = N_{switch}/N_{conc.}$. We also estimate the error in this estimate of p, by a simple binomial law, $err = \sqrt{p(1-p)/N_{conc.}}$. The success rates in Figs. 6, 8 are reported as $(p \pm err) \times 100\%$.

I. SHAPE protection scores are in agreement with consensus secondary structure

Sequence homologs in the RF00162 family are collected based on similarity to a group of manually curated sequences in the seed. Overall, for many of these sequences (both in the seed and in the full alignment), direct experimental evidence of their actual behavior and structure is limited, except for specific cases, such as the *Thermoanaerobacter teng- congensis* and the *Bacillus subtilis yitJ* SAM riboswitches, which have been extensively studied in the literature fueled by detailed knowledge of their published crystalized structures [41, 49]. For many

Cluster	Sites
Pseudoknot	25, 26, 27, 28, 77, 79
Kink-turn	34, 35, 36, 37
Base-triple	73, 74, 76, 100
SAM contact	46, 47, 102, 103
P1	101, 104, 105
Other sites	75

TABLE I. Hallmark sites of structural switch in response to SAM. List of sites that exhibit observable SHAPE reactivity differences upon SAM binding. Positions are numbered following the Rfam reference alignment. See also Fig. 9.

other sequences in the MSA, their actual behavior is at most hypothesized based on indirect evidence.

We have here obtained detailed SHAPE data of 151 sequences of the seed alignment. Our data shows that, in average, these sequences are compatible with the consensus secondary structure of the RF00162 family, shown in Fig. 1B. Indeed, we have computed the average protection scores $\langle \mathcal{S}(i) \rangle$ for each site i, over the sequences in the seed alignment probed in our experiments,

$$\langle \mathcal{S}(i) \rangle_{\text{seed}} = \frac{1}{B_{\text{seed}}} \sum_{n \in \text{seed}} \mathcal{S}_n(i) \tag{17}$$

Figure 9B plots $\langle \mathcal{S}(i) \rangle$ in the conditions with SAM and without SAM. Overall, the averaged protection scores are in detailed agreement with the consensus secondary structure of the aptamer, depicted in Fig. 9A. Helices P2, P3, P4 are seen to be base-paired in average in all conditions, with a mild overall increase in the values of $\mathcal{S}$ with the addition of magnesium and then SAM, indicating overal structural stabilization. The central junction loop (CL), and the loops on the second helix L2, the third helix L3, and the fourth helix L4, are consistently measured as reactive when SAM is not present, indicating that these sites are unpaired, as expected. Besides these major structural motifs, we also appreciate finer details such as the reactivity of single isolated bulge sites in positions 46 and 65 in absence of SAM. Next, comparing the behavior across different conditions, we appreciate the effect of magnesium and SAM on the structure. We highlight (in green) sites that change significantly in response to SAM. These include sites in direct contact with SAM (as known from the crystal structure [49]), and other tertiary motifs known to form in response to SAM. We discuss these next.

J. Selection of Hallmark sites

Based on these observations, we selected a number of hallmark sites across the aptamer sequence, for which we could rationalize observed reactivity changes in response to SAM binding, and which are consistent with expectations from previous chemical probing studies on SAM-I riboswitches. These sites are listed in I, and we discuss them next.

Pseudoknot. The pseudoknot is a tertiary contact formed between sites 24-28 on loop L2 and sites 77-80, along the junction between P3 and P4. Multiple sources of evidence point to the importance of this motif to SAM binding, from genetic studies [45], to crystal structure [41, 49], to sequence based statistical modelling [86]. Consistently with these previous observations, we find in our experiments that sites 24-28, 77 and 79 in natural sequences exhibit significant reactivity decrease upon SAM-binding, as can be appreciated in Fig. 5 and 9D. We therefore include these sites in Table I.

Sites 78, 80, also belonging to the pseudoknot, do not show significant protection upon SAM binding. As can be observed in the crystal structure (pdb 2GIS, [49]) site 80 at the edge of the pseudoknot is in a context and conformation favorable for the 1M7 probe to stack under the guanine and react with the cognate ribose, even when immobilized (E. Frezza, personal communication). This probably explains why site 80 remains slightly reactive even upon pseudoknot stabilization in the presence of the ligand. Site 78 is seen to exhibit protection in both conditions, likely due to other contacts formed outside the pseudoknot in absence of Mg^{2+}.

Base-triple. The A-minor motif helps create a groove where SAM is placed upon binding. It is an important ligand-dependent structural element [41, 49, 72]. While the two conserved G-C base pairs (21-30, 22-29) involved in this motif are stable and not reactive in any of the conditions assayed, our data clearly show consistent protections of A73 and 74.

The base-triple is an important tertiary motif observed in the bound structure of the aptamer [41, 49], involving nucleotides (24, 76 and 100) in between the A minor and the pseudoknot. Stabilization of this contact in response to SAM has been reported previously [29, 49]. Our data shows consistent protection of sites 76 and 101 in response to SAM, two of the sites involved in the base-triple.

Kink-turn. The kink-turn is a well characterized structural element of the SAM-riboswitch, with a central role in the fold of the aptamer, helping stabilize the coaxial stacking of the four helices and supporting the formation of the pseudoknot [72]. Stabilization of this tertiary structure in response to SAM has been observed, and evidenced both in the crystal structure [41, 49, 66], previous SHAPE experiments [29], and simulations [72]. Consistent with these previous observations, our data reveals significant protection at positions 34, 35, 36, 37 located in the kink-turn in response to both Mg^{2+} and SAM. We therefore include these sites in Table I.

SAM contacts. A number of sites are in direct contact with SAM in the bound structure, as has been established in published 3-dimensional structures [49]. SAM sits in a pocket between the interwoven P1 and P3 helices and the junction between P1 and P2, forming a network of contacts that results in stabilization of a number of sites. In

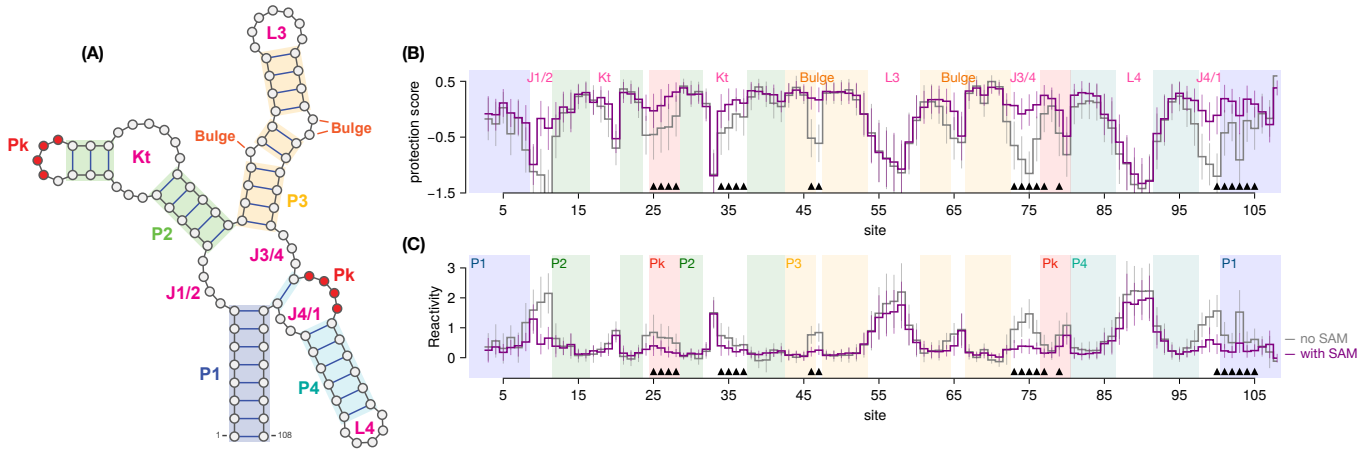


FIG. 9. **(A)** Annotated consensus secondary structure of the aptamer domain of the SAM-I riboswitch family (Rfam ID RF00162). **(B)** Average protection scores, $\langle \mathcal{S}(i) \rangle$ (see Eq. (16)) per site, of the natural probed sequences, for the two conditions: with SAM and no SAM. Error bars (standard deviation) are also shown. Hallmark sites (Table I) are indicated with black triangles. **(C)** Average site reactivities.

particular sites 46, 47, belonging to a bulge in P3, directly contact SAM. Since these sites are initially unpaired, we expect a reactivity decrease upon SAM binding. Sites 102, 103 of the 5'-end arm of the P1 helix are similarly embracing SAM. Since even in the isolated aptamer domain, P1 might be disordered in absense of SAM (see Fig. 1A), we expect significant reactivity decreases in this region as well upon SAM binding. Also sites 10-11 show response in the averaged reactivity profiles (Fig. 5), but for many individual sequences we had few data at these sites and therefore were not included.

P1 & other sites. Stabilization of the P1 helix in response to SAM has a key regulatory role, releasing a complementary sequence that forms the hairpin loop and blocks downstream transcription, see Fig. 1B. We observe reactivity changes in sites 101, 104 and 105 of P1, likely reflecting the SAM induced stabilization of P1. These sites are also near SAM in the bound structure, though not in direct contact [49].

Finally, we observe significant protection in site 75 upon SAM binding. It is flanked by sites participating in the A-minor (74) and base-triple (76) motifs, both of which are significantly protected in response to SAM. The protection of neighboring sites is likely to promote low reactivity at 75.

K. Principal component analysis

We carried out a principal component analysis (PCA) of the natural MSA. First, we one-hot encode the natural sequences in a $q \times N \times B$ binary tensor $\mathcal{D}$, where $B = 6161$ is the number of sequences in the full MSA collected above. The tensor has $\mathcal{D}^a_{in} = 1$ if sequence n of the alignment has symbol $a \in \{1, ..., 5\}$ at position i, and otherwise $\mathcal{D}^a_{in} = 0$. We then compute a covariance tensor, defined as follows

$$C^{ab}_{ij} = \frac{1}{B} \sum_n \mathcal{D}^a_{in} D^b_{jn} - \left(\frac{1}{B} \sum_n \mathcal{D}^a_{in} \right) \left(\frac{1}{B} \sum_n \mathcal{D}^b_{jn} \right) \tag{18}$$

We flatten the tensor C^{ab}_{ij} into a $qN \times qN$ matrix, and then perform a standard eigenvalue decomposition on it. Individual sequences are then projected along the two top components (with largest eigenvalue) of the decomposition.

L. Accompanying software

All the code necessary to reproduce the results in this paper is available as an open-source Julia [2, 12] package. It can be obtained from Github, at the following URL: `https://github.com/cossio/SamApp2024`.

V. ACKNOWLEDGEMENTS

We gratefully acknowledge Sean R. Eddy and Eric P. Nawrocki, for helpful discussions about Infernal. This work is supported by Grant No. ANR-19 Decrypted CE30-0021-01. J.F.d.C.D. acknowledges support from Université PSL, through the AI Junior Fellow program.Research in B.S laboratory is also funded by the CNRS, and Université Paris Cité. PH and FXLdM were supported by ANR DECRYPTED (ANR 19-CE30-0021-03) as well as ANR PRE-RIBO60S (ANR 20-CE12-0026-02); ANR PARNASSUS (ANR 19-CE45-0023-02) and as well as ANR INSANNE (ANR 21-CE45-0034-03) respectively.

Conflict of interest statement. None declared.

[1] R. T. Batey. Recognition of S-adenosylmethionine by riboswitches. *Wiley Interdisciplinary Reviews: RNA*, 2(2):299–311, 2011.

[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[3] A. Boussebayle, D. Torka, S. Ollivaud, J. Braun, C. Bofill-Bosch, M. Dombrowski, F. Groher, K. Hamacher, and B. Suess. Next-level riboswitch development—implementation of capture-selex facilitates identification of a new synthetic riboswitch. *Nucleic acids research*, 47(9):4883–4895, 2019.

[4] B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A. M. Walczak, T. Mora, S. Cocco, and R. Monasson. A transfer-learning approach to predict antigen immunogenicity and t-cell receptor specificity. *ELife*, 12:e85126, 2023.

[5] B. Bravi, J. Tubiana, S. Cocco, R. Monasson, T. Mora, and A. M. Walczak. RBM-MHC: A semi-supervised machine-learning method for sample-specific prediction of antigen presentation by HLA-I alleles. *Cell systems*, 12(2):195–202, 2021.

[6] S. Busan and K. M. Weeks. Accurate detection of chemical modifications in rna by mutational profiling (map) with shapemapper 2. *Rna*, 24(2):143–148, 2018.

[7] F. Calvanese, C. N. Lambert, P. Nghe, F. Zamponi, and M. Weigt. Towards parsimonious generative modeling of rna families. *arXiv preprint arXiv:2310.12700*, 2023.

[8] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, et al. The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC bioinformatics*, 3:1–31, 2002.

[9] J. Chappell, M. K. Takahashi, and J. B. Lucks. Creating small transcription activating rnas. *Nature chemical biology*, 11(3):214–220, 2015.

[10] S. Cocco, A. De Martino, A. Pagnani, and M. Weigt. Statistical-physics approaches to rna molecules, families and networks. *arXiv preprint arXiv:2207.13402*, 2022.

[11] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[12] S. Danisch and J. Krumbiegel. Makie.jl: Flexible high-performance data visualization for julia. *Journal of Open Source Software*, 6(65):3349, 2021.

[13] K. Darty, A. Denise, and Y. Ponty. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25(15):1974, 2009.

[14] G. De Bisschop, D. Allouche, E. Frezza, B. Masquida, Y. Ponty, S. Will, and B. Sargueil. Progress toward shape constrained computational prediction of tertiary interactions in rna structure. *Non-coding RNA*, 7(4):71, 2021.

[15] E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt. Direct-coupling analysis of nucleotide coevolution facilitates rna secondary and tertiary structure prediction. *Nucleic acids research*, 43(21):10444–10455, 2015.

[16] J. Deforges, N. Chamond, and B. Sargueil. Structural investigation of hiv-1 genomic rna dimerization process reveals a role for the major splice-site donor stem loop. *Biochimie*, 94(7):1481–1489, 2012.

[17] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate shape-directed rna structure determination. *Proceedings of the National Academy of Sciences*, 106(1):97–102, 2009.

[18] G. Desjardins, H. Luo, A. Courville, and Y. Bengio. Deep tempering. *arXiv preprint arXiv:1410.0123*, 2014.

[19] A. Di Gioacchino, J. Procyk, M. Molari, J. S. Schreck, Y. Zhou, Y. Liu, R. Monasson, S. Cocco, and P. Šulc. Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLoS computational biology*, 18(9):e1010561, 2022.

[20] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[21] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[22] S. R. Eddy. Computational analysis of conserved rna secondary structure in transcriptomes and genomes. *Annual review of biophysics*, 43:433–456, 2014.

[23] J. Fernandez-de Cossio-Diaz, S. Cocco, and R. Monasson. Disentangling representations in restricted boltzmann machines without adversaries. *Physical Review X*, 13:021003, Apr 2023.

[24] J. Fernandez-de-Cossio-Diaz, C. Roussel, S. Cocco, and R. Monasson. Accelerated sampling with stacked restricted boltzmann machines. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multistable rna molecules. *Rna*, 7(2):254–265, 2001.

[26] E. Frezza, A. Courban, D. Allouche, B. Sargueil, and S. Pasquali. The interplay between molecular flexibility and rna chemical probing reactivities analyzed at the nucleotide level via an extensive molecular dynamics study. *Methods*, 162:108–127, 2019.

[27] F. J. Grundy and T. M. Henkin. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Molecular microbiology*, 30(4):737–749, 1998.

[28] R. R. Gutell, J. C. Lee, and J. J. Cannone. The accuracy of ribosomal rna comparative structure models. *Current opinion in structural biology*, 12(3):301–310, 2002.

[29] S. P. Hennelly, I. V. Novikova, and K. Y. Sanbonmatsu. The expression platform and the aptamer: cooperativity between mg2+ and ligand in the sam-i riboswitch. *Nucleic acids research*, 41(3):1922–1935, 2013.

[30] 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(6):384–392, 2011.

[31] G. E. Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

[32] J. Hoetzel and B. Suess. Structural changes in aptamers are essential for synthetic riboswitch engineering. *Journal*

of Molecular Biology, 434(18):167631, 2022.

[33] C. Jeancolas, Y. J. Matsubara, M. Vybornyi, C. N. Lambert, A. Blokhuis, T. Alline, A. D. Griffiths, S. Ameta, S. Krishna, and P. Nghe. Rna diversification by a self-reproducing ribozyme revealed by deep sequencing and kinetic modelling. *Chemical Communications*, 57(61):7517–7520, 2021.

[34] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[35] I. Kalvari, E. P. Nawrocki, J. Argasinska, N. Quinones-Olvera, R. D. Finn, A. Bateman, and A. I. Petrov. Non-coding rna analysis using the rfam database. *Current protocols in bioinformatics*, 62(1):e51, 2018.

[36] I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Research*, 49(D1):D192–D200, 2021.

[37] K. Kavita and R. R. Breaker. Discovering riboswitches: the past and the future. *Trends in biochemical sciences*, 48(2):119–141, 2023.

[38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Gaikwad, S. Yoon, A. Treuille, et al. Rna design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6):2122–2127, 2014.

[40] R. Lorenz, S. H. Bernhart, C. Hönerzu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14, 2011.

[41] C. Lu, F. Ding, A. Chowdhury, V. Pradhan, J. Tomsic, W. M. Holmes, T. M. Henkin, and A. Ke. SAM recognition and conformational switching mechanism in the Bacillus subtilis yitJ S box/SAM-I riboswitch. *Journal of molecular biology*, 404(5):803–818, 2010.

[42] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[43] C. Malbranke, W. Rostain, F. Depardieu, S. Cocco, R. Monasson, and D. Bikard. Computational design of novel cas9 pam-interacting domains using evolution-based modelling and structural quality assessment. *PLOS Computational Biology*, 19(11):e1011621, 2023.

[44] M. Mandal, M. Lee, J. E. Barrick, Z. Weinberg, G. M. Emilsson, W. L. Ruzzo, and R. R. Breaker. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, 306(5694):275–279, 2004.

[45] B. A. McDaniel, F. J. Grundy, and T. M. Henkin. A tertiary structural element in s box leader rnas is required for s-adenosylmethionine-directed transcription termination. *Molecular microbiology*, 57(4):1008–1021, 2005.

[46] J. Melchior, A. Fischer, and L. Wiskott. How to center deep boltzmann machines. *The Journal of Machine Learning Research*, 17(1):3387–3447, 2016.

[47] P. Meysman, J. Barton, B. Bravi, L. Cohen-Lavi, V. Karnaukhov, E. Lilleskov, A. Montemurro, M. Nielsen, T. Mora, P. Pereira, A. Postovskaya, M. Rodriguez Martinez, J. Fernandez-de-Cossio-Diaz, A. Vujkovic, A. M. Walczak, A. Weber, R. Yin, R. Eugster, and V. Sharma. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics*, 9:100024, 2023.

[48] F. Michel and E. Westhof. Modelling of the three-dimensional architecture of group i catalytic introns based on comparative sequence analysis. *Journal of molecular biology*, 216(3):585–610, 1990.

[49] R. K. Montange and R. T. Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, 441(7097):1172–1175, 2006.

[50] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[51] E. P. Nawrocki and S. R. Eddy. *INFERNAL User's Guide: Sequence analysis using profiles of RNA sequence and secondary structure consensus*. INFERNAL development team.

[52] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[53] P. Nawrocki Eric and R. I. Eddy Sean. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[54] R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

[55] R. Pearce, G. S. Omenn, and Y. Zhang. De novo rna tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv*, pages 2022–05, 2022.

[56] I. R. Price, J. C. Grigg, and A. Ke. Common themes and differences in SAM recognition among SAM riboswitches. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(10):931–938, 2014.

[57] X. Qiu. Sequence similarity governs generalizability of de novo deep learning models for rna secondary structure prediction. *PLOS Computational Biology*, 19(4):e1011047, 2023.

[58] E. Rivas. Rna structure prediction using positive and negative evolutionary information. *PLoS computational biology*, 16(10):e1008387, 2020.

[59] E. Rivas, R. Lang, and S. R. Eddy. A range of complex probabilistic models for rna secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18(2):193–212, 2012.

[60] O. Rivoire. How flexibility can enhance catalysis. *Physical Review Letters*, 131(8):088401, 2023.

[61] D. A. Rodionov, I. Dubchak, A. Arkin, E. Alm, and M. S. Gelfand. Reconstruction of regulatory and metabolic pathways in metal-reducing δ-proteobacteria. *Genome biology*, 5(11):1–27, 2004.

[62] S. Roy, J. N. Onuchic, and K. Y. Sanbonmatsu. Cooperation between magnesium and metabolite controls collapse of the sam-i riboswitch. *Biophysical journal*, 113(2):348–359, 2017.

[63] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

[64] A. Saaidi, D. Allouche, M. Regnier, B. Sargueil, and Y. Ponty. IPANEMAP: integrative probing analysis of nucleic acids empowered by multiple accessibility profiles. *Nucleic acids research*, 48(15):8276–8289, 2020.

[65] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

[66] K. T. Schroeder, P. Daldrop, and D. M. Lilley. Rna tertiary interactions in a riboswitch stabilize the structure of a kink turn. *Structure*, 19(9):1233–1240, 2011.

[67] V. Sharma, Y. Nomura, and Y. Yokobayashi. Engineering complex riboswitch regulation by dual genetic selection. *Journal of the American Chemical Society*, 130(48):16310–16315, 2008.

[68] K. Shimagaki and M. Weigt. Selection of sequence motifs and generative hopfield-potts models for protein families. *Physical Review E*, 100(3):032128, 2019.

[69] 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(9):959–965, 2014.

[70] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[71] C. D. Stoddard and R. T. Batey. Mix-and-match riboswitches, 2006.

[72] C. D. Stoddard, R. K. Montange, S. P. Hennelly, R. P. Rambo, K. Y. Sanbonmatsu, and R. T. Batey. Free state conformational sampling of the SAM-I riboswitch aptamer domain. *Structure*, 18(7):787–797, 2010.

[73] N. Sudarsan, M. C. Hammond, K. F. Block, R. Welz, J. E. Barrick, A. Roth, and R. R. Breaker. Tandem riboswitch architectures exhibit complex gene control functions. *Science*, 314(5797):300–304, 2006.

[74] S. Sumi, M. Hamada, and H. Saito. Deep generative design of rna family sequences. *Nature Methods*, pages 1–9, 2024.

[75] M. Szikszai, M. Wise, A. Datta, M. Ward, and D. H. Mathews. Deep learning models for rna secondary structure prediction (probably) do not generalize across families. *Bioinformatics*, 38(16):3892–3899, 2022.

[76] A. Tareen and J. B. Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020.

[77] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

[78] R. J. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das, and R. O. Dror. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021.

[79] 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(18):6624–6629, 2014.

[80] J. Tubiana, S. Cocco, and R. Monasson. Learning protein constitutive motifs from sequence data. *Elife*, 8:e39397, 2019.

[81] D. H. Turner and D. H. Mathews. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, 38(suppl_1):D280–D282, 2010.

[82] Q. Vicens and J. S. Kieft. Thoughts on how to think (and talk) about rna structure. *Proceedings of the National Academy of Sciences*, 119(17):e2112677119, 2022.

[83] A. M. Watkins, R. Rangan, and R. Das. Farfar2: improved de novo rosetta prediction of complex global rna folds. *Structure*, 28(8):963–976, 2020.

[84] J. E. Weigand, S. R. Schmidtke, T. J. Will, E. Duchardt-Ferner, C. Hammann, J. Wöhnert, and B. Suess. Mechanistic insights into an engineered riboswitch: a switching element which confers riboswitch activity. *Nucleic acids research*, 39(8):3363–3372, 2011.

[85] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[86] C. Weinreb, A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander, and D. S. Marks. 3d rna and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975, 2016.

[87] H.-T. Yao, Y. Ponty, and S. Will. Developing complex RNA design applications in the Infrared framework. In *RNA Folding - Methods and Protocols*. 2022.

[88] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of mathematical biology*, 46:591–621, 1984.