Designing molecular RNA switches with Restricted Boltzmann machines

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

Published online DD MM YYYY *Nucleic Acids Research, YYYY, Vol. xx, No. xx* **1–21**
doi:10.1093/nar/gkn000

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

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

Riboswitches are structured allosteric RNA molecules capable of switching between competing conformations in response to a metabolite binding event, eventually triggering a regulatory response. Computational modelling of these molecules is complicated by complex tertiary contacts, conditioned to the presence of their cognate metabolite. In this work, we show that Restricted Boltzmann machines (RBM), a simple two-layer machine learning model, capture intricate sequence dependencies induced by secondary and tertiary structure, as well as the switching mechanism, resulting in a model that can be successfully used for the design of allosteric RNA. As a case study we consider the aptamer domain of SAM-I riboswitches. To validate the functionality of designed sequences experimentally by SHAPE-MaP, we develop a tailored analysis pipeline adequate for high-throughput probing of diverse homologous sequences. We find that among the probed 84 RBM designed sequences, showing up to 20% divergence from any natural sequence, about 28% (and 47% of the 45 among them having low RBM effective energies), are correctly structured and undergo a structural allosteric in response to SAM. Finally, we show how the flexibility of the molecule to switch conformations is connected to fine energetic features of its structural components.

INTRODUCTION

Riboswitches are regulatory RNA elements commonly found in bacterial messenger RNAs (mRNAs) upstream of the coding sequence, capable of binding specific cellular metabolites and inhibit or shutdown the expression of downstream genes at either the transcriptional or translational level (1, 68). One the largest studied groups are the SAM-riboswitches (50), initially discovered as a conserved motif present upstream of a number of genes involved in sulfate assimilation into cysteine and methionine biosynthesis in bacteria (23). They recognize S-adenosyl methionine (SAM) as their effector metabolite and consist of two domains: an *aptamer* part, which is specific to SAM binding, and the *expression platform*. In order to perform their function, SAM-riboswitches can switch between two competing structural folds in response to the binding of their cognate metabolite. Most SAM riboswitches characterized to date have been shown to regulate at a transcriptional level, as follows (19, 20, 76, 80):

- In absence of SAM, the 3'-end of the aptamer sub-domain captures a complementary sub-sequence in the expression platform, which is then unable to form the terminator hairpin loop. This conformation is compatible with continuation of transcription, which results in the eventual expression of a downstream gene. This is the ON state, depicted in Figure 1A.

- In presence of SAM, which binds within a pocket in the aptamer sub-domain, the expression platform forms a hairpin loop, which acts as a terminator of transcription by recruiting NusA and prompting early release of the nascent transcript from the RNA polymerase. Therefore, the downstream gene is not expressed. This is the OFF state, depicted in Figure 1B.

While six different SAM binding structural motifs have been identified in nature, this study focuses on type I SAM aptamers (SAM-I) (1). Some SAM-I aptamers regulate gene expression at a translational level, by sequestering a downstream Shine-Dalgarno sequence instead of forming the terminator hairpin-loop (68). For simplicity, we focus on transcriptional regulation in our description, since this has been the prevalent mechanism studied in the literature to date.

The conformational flexibility needed for riboswitch function is facilitated by several tertiary contacts that cooperate to create a pocket around SAM (43, 50). Figure

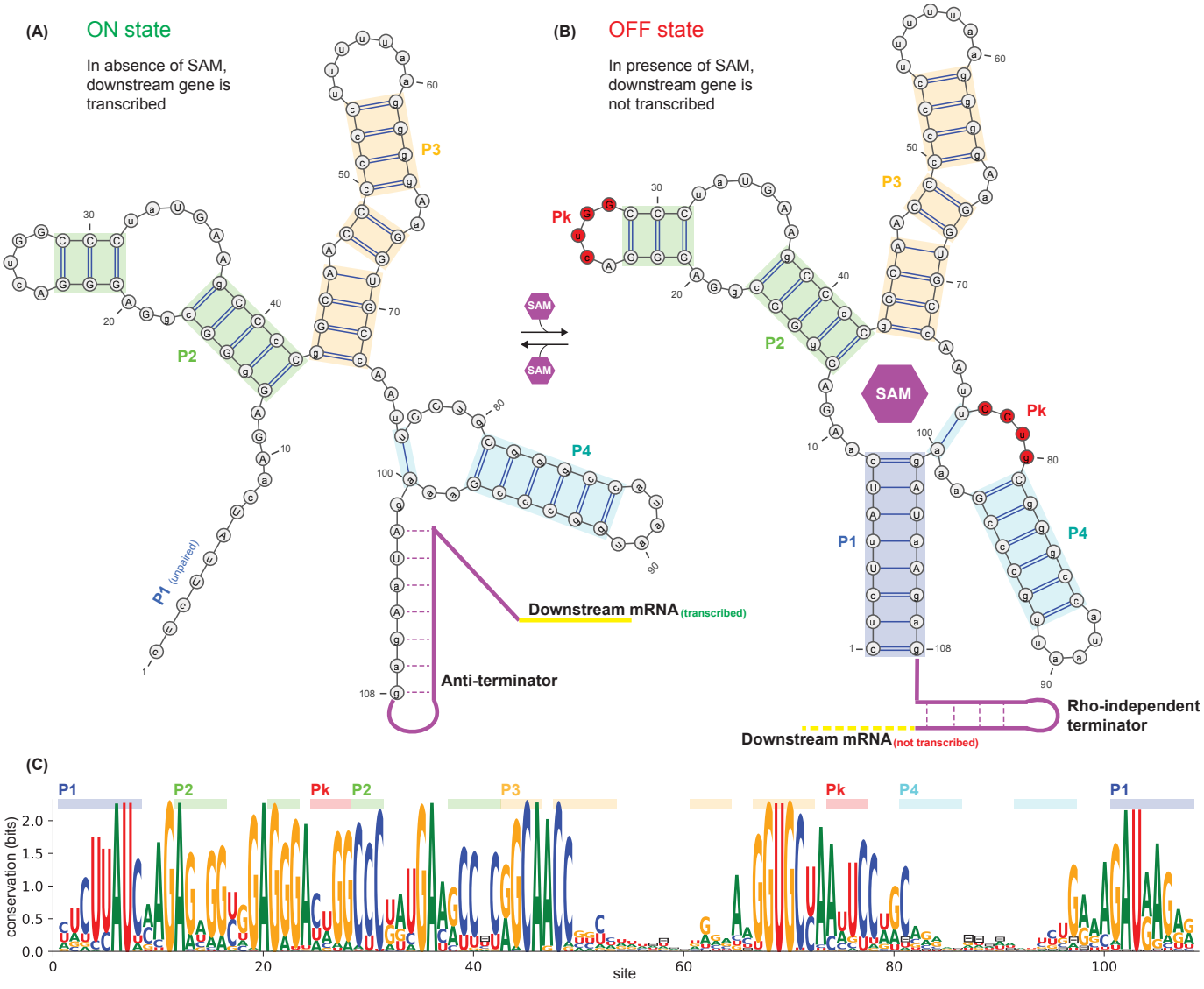

Figure 1. Structure, regulatory function, and sequence conservation of the aptamer domain of the SAM-I riboswitch, acting at a transcription level. **(A)** In absence of SAM, the P1 helix of the aptamer domain is unpaired, leaving the 3'-end free to pair with the anti-terminator segment of the expression platform. This conformation is incompatible with formation of a terminator motif, resulting in transcription of the downstream gene. **(B)** SAM (represented by the purple hexagon) is captured in a groove contacting several sites around the central four-way junction. In the bound 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. 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. **(C)** Sequence conservation logo of aligned homologs of the SAM-I riboswitch aptamer domain family (RF00162 on Rfam). Gaps are indicated by an empty-set character (∅). Secondary structure plots obtained with VARNA (10).

1A shows the secondary structure of the aptamer domain in absence of SAM, where transcription is allowed (the ON state), while panel B depicts the structure when SAM is bound and transcription continuation is prevented (the OFF state). The downstream expression platform domain and following mRNA gene segments are also represented schematically in each case. SAM binding is reported to stabilize the formation of a pseudoknot in the aptamer structure (50), indicated in red in the figure. The presence of this pseudoknot has been validated genetically (39) to be essential for riboswitch function, and is directly observed in crystal structures of the SAM-bound aptamer (43).

Because of the "mix-and-match" nature of riboswitch aptamers and expression platforms (64), many authors chose to study an isolated aptamer domain, outside the context of its host RNA in the hope of yielding insights into the general behavior of ligand recognition (65) and the structural switch in response to ligand binding. As currently understood, the aptamer domain exerts control over the surrounding sequence (rather than the converse), and thus the regulatory outcome is a function of ligand binding, which can be understood by looking at the isolated aptamer. In support of this modularity, full riboswitches or aptamer domains have been found to exist in tandem, with specifities to the same or different ligands and

functioning independently, resulting in sophisticated Boolean logic-like regulatory gates, that permit expression of the downstream gene only when one or more ligands are present (38, 54, 64, 66). Understanding how the aptamer domain by itself is able to implement a structural switch in response to the ligand is an important first step in the direction of understanding the mechanisms of the full riboswitch.

In general, 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 RNA sequences sampled during evolution and collected in Multiple Sequence Alignments (MSA) (46) have been applied to predict RNA structures from sequence only (5, 42, 53). Computational methods exploiting sequence co-variation have been successful in predicting RNA secondary structure, even before experimental probing (18, 24). The majority of the computational algorithms introduced to predict the structure of RNAs rely on strong simplifying assumptions that are enforced for purely technical reasons (mainly computational efficiency (53, 75)), such as ignoring pseudo-knots and other tertiary contacts, which however are known to be biologically important. SAM-I riboswitches, in particular, form a complex network of tertiary interactions around SAM during binding, facilitating the structural switch (43), but posing difficulties to prevalent secondary structure modelling tools (35). Covariation analysis has been pursued with the aim of predicting also pseudoknots and other tertiary contacts from statistical couplings inferred from the conservation and covariations in the MSA columns (12, 79), or by including positive and negative evolutionary information such as in the Cascade covariation Folding Algorithm (CacoFold) (52). Purely machine learning approaches have recently shown promising results in RNA structure prediction. ARES, a geometric deep learning approach that attempts to rank candidate structures for a sequence by their closeness to the (possibly unknown) true crystal structure, (71), can be used as an aid to sampling methods such as Rosetta FARFAR2 (77) which typically produce a large number of decoys without any indication of which one is more correct. DeepFoldRNA (49), using techniques similar to AlphaFold (29), significantly outperformed the state-of-the-art in tertiary structure prediction from sequence alone for common RNA molecule benchmarks. Although these approaches look promising, it is still an open question to understand why AlphaFold-level accuracies are not reached in RNA structure prediction. The mirroring problem of sequence design, folding in a particular structure or performing a desired function, has also long been investigated, especially focusing on adopting target minimum free energy (MFE) secondary structures by rational design (22, 81) . The issue of building generative models effective in RNA design, not only with structural but more generally a functional target, is still an outstanding problem with many potential applications.

To address some of these questions, we employ in this work Restricted Boltzmann machines (RBM), a simple two-layer neural network. RBM can be regarded as building blocks of deeper neural architectures (15, 27, 55), have been successfully applied to diverse machine learning tasks (21, 58), and more recently to modelling protein sequences in various contexts (3, 4, 41, 73). We use RBM to develop a probabilistic model of sequences in the SAM-I riboswitch family. As we will show, RBM capture constraints acting at the sequence level that enable aptamers to adopt the correct secondary structure, form tertiary contacts and moreover to effect a conformational switch in response to SAM, compatible with the role of the aptamer domain in the regulatory function of riboswitches.

This paper is structured as follows. In Methods, we describe our pipeline, going from sequence data acquisition from Rfam (31), to our implementation and training of the RBM, and finally the experimental validation of designed sequences by SHAPE-MaP probing (11, 14, 67). Then in Results, we study the features and constraints extracted by the RBM, and evaluate experimentally the structural response of generated and natural aptamers to SAM binding.

MATERIALS AND METHODS

Multiple sequence alignment of SAM-I riboswitches

The RF00162 family from the Rfam database (31) 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 purported aptamer sequences supported by literature evidence. These seed sequences are aligned to a consensus secondary structure (shown in Figure 1B) that has been informed by the holo-form of SAM-I riboswitch crystal structures (36, 43). 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) (17) on this seed alignment using Infernal (46) with default settings. Following standard protocols (30), 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 Figure 1A.

Infernal pipeline

Infernal (46) 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) (17). 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 (30, 31). We employed Infernal to construct the RF00162 full MSA that we use to train the RBM with the procedure described in the previous section, as explained above.

Fundamental assumptions at the core of the CM enable implementation of efficient dynamical programming algorithms to train, sample, and scan large genomes for sequences that score highly under the CM (17). 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.

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 (45). We will refer to this CM model as the *Rfam CM*.

Refined CM. Since the *Rfam CM* is strongly regularized, in this work, we also trained another CM model on the full MSA, with no regularization, that we call hereafter the *Refined 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.

Untangled 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 (Figure 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 "untangled", 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 *Untangled CM* in what follows.

Sampling the CMs. To better understand the limitations of CM models and the advantages of RBM, we sampled 10000 sequences from each of the three CMs described above. For the Untangled 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 the a given sequence. We computed this score using `cmalign`, with `-g` (global) option to avoid local approximations (45).

Restricted Boltzmann machines

Definitions. Restricted Boltzmann Machines (RBM) (27) 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 Figure 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,

$$Z=\sum_{\mathbf{v}}\int e^{-E(\mathbf{v},\mathbf{h})}d\mathbf{h} \tag{2}$$

and the energy $E(\mathbf{v},\mathbf{h})$ is given by

$$E(\mathbf{v},\mathbf{h})=\sum_{i=1}^{N}\mathcal{V}_i(v_i)+\sum_{\mu=1}^{M}\mathcal{U}_\mu(h_\mu)-\sum_{i=1}^{N}\sum_{\mu=1}^{M}w_{i\mu}(v_i)h_\mu \tag{3}$$

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 $\mathcal{V}_i(v_i)$, 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 (73),

$$\mathcal{U}_\mu(h_\mu)=\begin{cases}\gamma_\mu^+h_\mu^2/2-\theta_\mu^+h_\mu & h_\mu\geq 0\\ \gamma_\mu^-h_\mu^2/2-\theta_\mu^-h_\mu & h_\mu\leq 0\end{cases} \tag{4}$$

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 $\theta_\mu^+=\theta_\mu^-$, (**4**) becomes equivalent to a single quadratic (*i.e.*, Gaussian) potential, and is closely related to Direct-Coupling Analysis models popular in protein sequence modelling (8, 12, 44, 56, 61, 78). 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 (7, 82)), as well as functional interactions that can span complex structural and sequence motifs. dReLU can also adopt a bimodal form when $\theta_\mu^+\gg 1$ and $\theta_\mu^-\ll -1$, 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:

$$P(\mathbf{v})=\frac{1}{Z}\int e^{-E(\mathbf{v},\mathbf{h})}d\mathbf{h}=\frac{1}{Z}e^{-E_{\text{eff}}(\mathbf{v})} \tag{5}$$

where $E_{\text{eff}}(\mathbf{v})$ is the resulting energy as a function of visible configurations $\mathbf{v}$ only, that incorporates effective interactions

arising from the marginalized latent variables (see Figure 2C):

$$E_{\text{eff}}(\mathbf{v}) = \sum_i \mathcal{V}_i(v_i) - \sum_\mu \ln \int e^{\sum_i w_{i\mu} v_i h_\mu - \mathcal{U}_\mu(h_\mu)} dh_\mu \quad \textbf{(6)}$$

Although evaluating $P(\mathbf{v})$ is computationally difficult (because the partition function Z is intractable), Equation (**6**) shows that the effective energy $E_{\text{eff}}(\mathbf{v})$ can be computed in linear time in the number of units of the RBM.

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 by maximizing the average log-likelihood of the data

$$\mathcal{L} = \frac{1}{B_{\text{MSA}}} \sum_{\mathbf{v} \in \text{MSA}} \log P(\mathbf{v}), \quad \textbf{(7)}$$

plus a regularization term,

$$\mathcal{R} = -\frac{\lambda_{\text{reg}}}{2} \sum_\mu \left(\sum_i |w_{i\mu}| \right)^2, \quad \textbf{(8)}$$

where the sum is taken over the sequences in the MSA, which consists of B_{MSA} sequences, and λ_{reg} is a non-negative regularization parameter. This form of the regularization, combining L_2 and L_1 norms, has been proposed by (73) 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 (27).

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:

$$\omega_{t+1} = \omega_t + \eta \frac{\partial}{\partial \omega} (\mathcal{L} + \mathcal{R}), \quad \textbf{(9)}$$

where η is a small positive learning rate. In practice, the optimization is accelerated by an adaptive momentum term (32) and a centering trick (40), following the implementation of RBM in (21). 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 (27). We employ the Persistent Contrastive Divergence (PCD) algorithm (70), where a number of Markov chains sampled from the model are updated in each parameter update. We have that:

$$\frac{\partial \mathcal{L}}{\partial \omega} = \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{MSA}}}_{\text{positive gradient}} - \underbrace{\left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{RBM}}}_{\text{negative gradient}}. \quad \textbf{(10)}$$

The first term is an empirical average performed over the data, as in (**7**), while the second term is averaged over sequences sampled from the RBM. We represent these terms as arrows in Figure 2B. The first term (blue), tends to drive the parameters ω of the RBM such that the effective energy $E_{\text{eff}}(\mathbf{v})$ of sequences $\mathbf{v}$ in the data is lowered. 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 mass implies that regions of sequence space not populated by data sequences must be penalized. This is taken care of by the second term in (**10**) (in red), which tends to increase the average energy of sequences uniformly sampled from the RBM. This allows the RBM to uncover constraints from the data, such as complementarity of base-pairs in an helix of the secondary structure or avoidance of base-pairs in a non-functional competing fold. Violation of these constraints is likely to result in loss of function. The net effect of the two terms (also called positive and negative phase terms in earlier papers (27, 58)), 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 Figure 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.

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:

$$\begin{aligned} P(\mathbf{v}|\mathbf{h}) &\propto \prod_i \exp\left(-\mathcal{V}_i(v_i) + \sum_\mu w_{i\mu}(v_i) h_\mu \right) \\ P(\mathbf{h}|\mathbf{v}) &\propto \prod_\mu \exp\left(-\mathcal{U}_\mu(h_\mu) + \sum_i w_{i\mu}(v_i) h_\mu \right) \end{aligned} \quad \textbf{(11)}$$

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 of the machine 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 RBMs we trained in this work, we found that $\sim$1000 Gibbs sampling steps were more than sufficient to reach equilibirum.

Estimating the partition function. Evaluating the likelihood $P(\mathbf{v})$ of a given test sequence $\mathbf{v}$ requires computing the partition function Z in (**2**), which involves an untractable summation over all the possible q^N sequences. An approximate Monte-Carlo scheme known as annealed

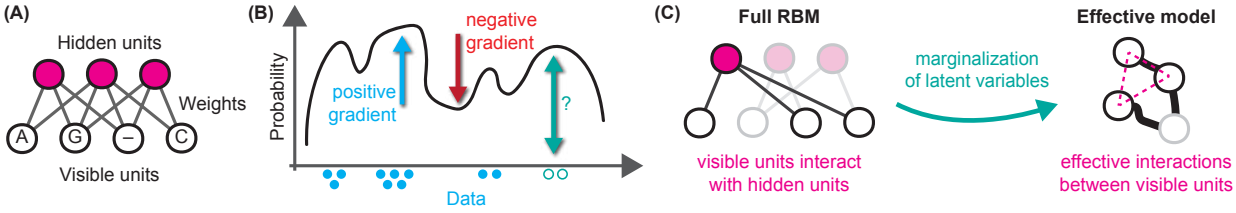

Figure 2. Restricted Boltzmann machines. **(A)** An RBM consists of two layers: visible and hidden. Each visible unit represents a site in an aligned sequence, and takes values A, C, G, U, (corresponding to the four nucleotides) or – (the alignment gap symbol). Hidden units represent features, automatically extracted from the data during learning. The two layers are connected by parameters, called *weights*. No connections are allowed within units of the same layer. **(B)** The RBM is trained by maximization of a regularized likelihood, see equation (**10**). A gradient term moves probability mass towards regions in sequence space densely populated by data, automatically discovering features desirable for functional sequences (blue). An opposite gradient term removes probability mass from regions unoccupied by data, automatically discovering constraints that if violated result in non-functional sequences (red). After training, the RBM also assigns positive probability to interesting regions not covered by data (teal). **(C)** The RBM is able to model complex interactions along the RNA sequence. On the left, a hidden unit interacting with three visible units is highligited. After marginalizing over the hidden units configurations, effective interactions arise between the visible units, see Equation (**6**). Here we represent schematically a three-body interaction, arising from the three connections of the summed hidden unit.

importance sampling allows us to obtain estimates of Z (48). Our implementation follows (21).

Further details. Our implementation of the RBM training, sampling and partition function calculations, follows closely that of (21). Additional details can be found in the Supplementary Note A.

Biophysical energy calculations

We computed biophysical binding energy predictions for formation of P1 and the pseudoknot of various sequences using the Turner energy model, as implemented in the ViennaRNA package (35), with the RNAeval program. For the P1 helix, we computed the energy difference of each sequence in the consensus secondary structure, which has P1 paired (shown in Figure 1B), and in a conformation where P1 is unpaired (Figure 1A).

To estimate the binding energy associated to the pseudoknot, we used RNAeval on a secondary structure that includes only the pseudoknot base paired sites, with all other sites unpaired. We then considered only 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 with other structural elements (and other tertiary contacts) simultaneously.

SHAPE-MaP experiments

Selection of a representative set of natural SAM-I aptamers sequences. SAM-I aptamer natural sequences were downloaded from the RFAM database (31). Then CD-hit clustering program (`https://github.com/weizhongli/cdhit-web-server`) was used to select a set of representative sequences (34). An identity cut-off of 85% and 70% was applied for sequences extracted from the seed alignment and the full SAM-I sequence database respectively. In total, 206 clusters were generated with a representative sequence corresponding to each one (151 sequences from to the seed MSA and 55 sequences from to the full MSA).

RNA preparation. DNA oligonucleotides representing the 206 sequences of the natural aptamers and the 100 sequences of the artificial aptamers 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 biding site were purchased as an oligonucleotide pool (Twist bioscience®). The Tag sequence was designed to avoid interference with the aptamer secondary structure using RNAFold (35). The oligo pool was PCR amplified using the T7 promoter as forward primer and and five different reverse primers (see supplementary material). RNA was transcribed and prepared as previously described (13), and was checked for the absence of aberrant products by gel electrophoresis.

SHAPE probing and analysis. SHAPE chemical probing was performed as described previously (62). 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, 0.5 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®), a NGS library was prepared using NEBNext Ultra II DNA Library Prep Kit (New England Biolabs®) and final products were sequenced (NextSeq 500/500 Mid 2x150 flowcell). Sequencing data were analyzed and reactivity maps were derived using shapemapper. IPANEMAP (Saaidi et al. 2020) was used to generate RNA structure models for each sequences. In the end, the 306 selected sequences were probed in the following conditions:

- 30°C, without Mg^{2+} and without SAM (30°C)
- 30°C, with magnesium (Mg^{2+}).
- 30°C, with magnesium and 3 concentrations of SAM (SAM+Mg).

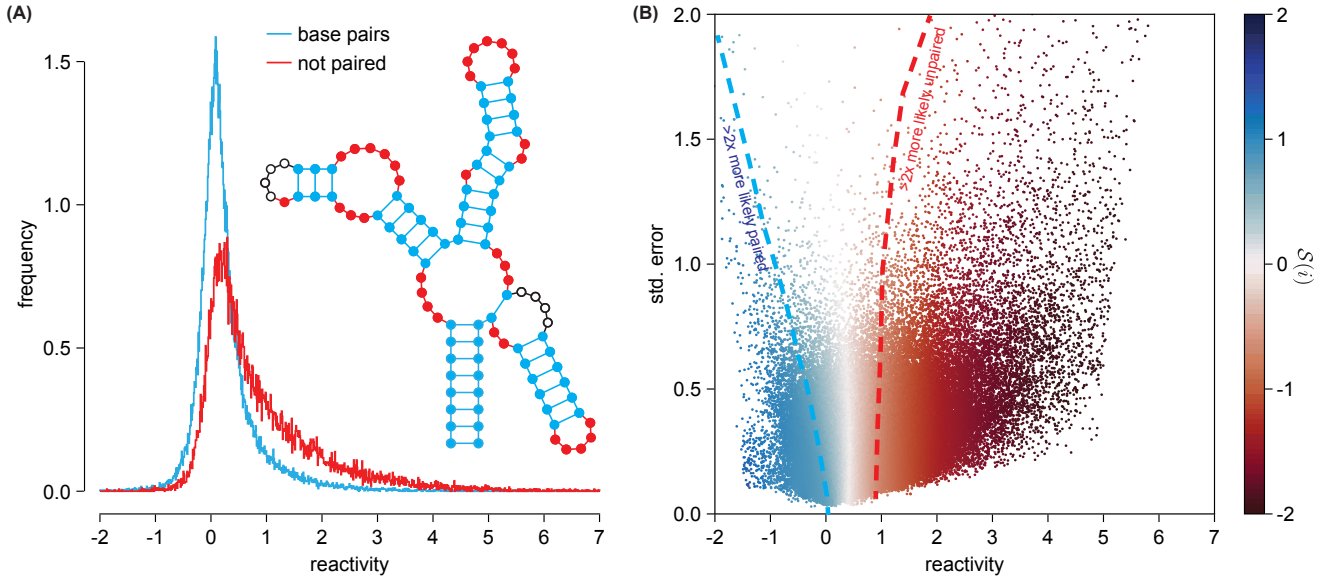

Figure 3. Statistical differences of SHAPE reactivities in paired and unpaired sites. **(A)** Histogram of reactivities of base-paired (blue) and unpaired sites (red) in probed natural sequences belonging to the manually curated seed alignment. The inset shows the consensus secondary structure with sites colored according to whether they are paired or unpaired. Sites with ambigous behavior (P1 helix, pseudoknot, indicated by empty circles in the inset secondary structure) are excluded from both histograms. **(B)** Scatter plot of measured reactivities (on the x-axis) and their estimated standard errors (on the y-axis), as estimated by the standard ShaPE-Mapper protocol (62), by first-order error propagation through the Poisson statistics of the mutation counts. The points are colored by the value of the log-odds ratio $\ln(P(\tilde{r}|\text{bp})/P(\tilde{r}|\text{np}))$ (see color bar), computed as explained in the text surrounding equation (**14**). The blue (red) dashed line indicates a contour separating sites over two times more likely to be paired (unpaired) than not.

Each probing reaction was repeated in triplicate.

Statistical analysis of SHAPE-MaP reactivities

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 (62):

$$r_{ni} = \frac{m_{ni} - u_{ni}}{d_{ni}} \qquad (\mathbf{12})$$

where m_{ni} is the mutation rate in presence of the reagent, u_{ni} is the mutation rate in its absence intended to cancel out sequencing error biases, and d_{ni} is the mutation rate in a de-naturating condition where the RNA is expected to be unfolded, intended to cancel site-dependent biases. Working with r_{ni} is usually better since this form is purported to cancel site-dependent biases in the raw SHAPE mutation rates, m_{ni}. The basis of the SHAPE-MaP pipeline relies on differences in the distribution of reactivities in base-paired and unpaired sites (62). To confirm these differences in our data, we considered the subset of probed sequences that belong to the manually curated sequences of the seed alignment, since these are the sequences for which we have more confidence in their compatibility with the consensus secondary structure of the SAM-I riboswitch family depicted in Figure 1B. We separated the $N = 108$ sites into base-paired or unpaired, excluding the 8 sites involved in the pseudoknot (see inset in Figure 3A). We then plotted the histograms of reactivities r_{ni} in Figure 3A. As expected, there is a robust statistical difference between the two sets of reactivities. Base-paired sites (blue in the figure) tend to have lower reactivities, consistent with the fact that they are less flexible in the RNA and hence less prone to react with the SHAPE reagent. On the contrary, unpaired sites (red in the figure) are more flexible and consequently more reactive to the SHAPE reagent, consistent with their statistically higher reactivities in Figure 3A.

Sampling noise. The finite number of sequencing reads collected at a site implies a statistical error in the reactivity computed by (**12**). Therefore, we cannot directly access the true reactivity r_{ni} at a site, but rather an experimental measurement $\tilde{r}_{ni}$ 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_{ni}, depends only on whether the site is base-paired or not. Intuitively, the form of the definition (**12**) is intended to (approximately) cancel other site-specific influences on the reactivity. Under this assumption, we can write:

$$\frac{P_{ni}(\tilde{r}_{ni}|\text{bp})}{P_{ni}(\tilde{r}_{ni}|\text{np})} = \frac{\int P(r|\text{bp}) P_{ni}(\tilde{r}_{ni}|r) dr}{\int P(r|\text{np}) P_{ni}(\tilde{r}_{ni}|r) dr} \qquad (\mathbf{13})$$

where:

- $P_{ni}(\tilde{r}_{ni}|\text{bp})$ is the probability of measuring a reactivity $\tilde{r}_{ni}$ 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.
- $P_{ni}(\tilde{r}_{ni}|r)$ is the probability of measuring reactivity $\tilde{r}_{ni}$ 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.
- $P_{ni}(\tilde{r}_{ni}|\text{np})$ 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 histograms shown in Figure 3A, under the (approximate) assumption that sequencing errors are averaged out when all the reactivity measurements are aggregated together. The kernel function used corresponds to a standard normal, with a bandwidth set according to the Silverman rule (63).

Applying Bayes theorem (37) in (**13**), we can write:

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

where $P(r)$ is the histogram of real reactivities, regardless of whether a site is paired or not,

$$\begin{aligned} P(r) &= P(r|\text{bp})P(\text{bp}) + P(r|\text{np})P(\text{np}) \\ &= \frac{P(r|\text{bp}) + P(r|\text{np})}{2} \end{aligned} \tag{15}$$

where we set $P(\text{bp}) = P(\text{np}) = 1/2$, as the most unbiased choice. The posterior $P_{ni}(r|\tilde{r}_{ni})$, on the other hand, quantifies our 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 (**12**). Since the mutation count at a site can be modeled by a Poisson distribution (62), the resulting form for the posterior distributions of the mutation rates m, u, d is a Gamma distribution, assuming a convenient choice of conjugate prior (37). Then, we can produce a Monte-Carlo estimate of $P_{ni}(r|\tilde{r}_{ni})$ by sampling of the posterior Gamma distributions of m, u, d, and then computing the reactivity through (**12**). 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 (**14**). In this way, we produce estimates of the ratios $P_{ni}(\tilde{r}_{ni}|\text{bp})/P_{ni}(\tilde{r}_{ni}|\text{np})$, quantifying the odds that a site is paired. Figure 3B shows a scatter plot of reactivities in our dataset, with the standard-error estimated by the standard SHAPE-Mapper pipeline (62) (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 (**14**). 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 (**14**), we are considering both the reactivity value and its uncertainty in assessing the plausibility that a site is paired or not.

Structural motifs and SHAPE evidence scores. We can exploit the odd-ratios $P_{ni}(\tilde{r}_{ni}|\text{bp})/P_{ni}(\tilde{r}_{ni}|\text{np})$ 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 an odds-ratio that motif $\mathcal{M}$ is present in sequence n, from the reactivity data, as follows:

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

which we take as the definition of the SHAPE scores $\mathcal{S}_n(\mathcal{M})$ 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.

To summarize, we define the scores $\mathcal{S}_n(\mathcal{M})$ to assess the presence of a base-paired motif $\mathcal{M}$. Similarly, we will denote by $\mathcal{S}_n(i)$ the log-odds ratio that a site i is base-paired in sequence n, (**14**), according to the SHAPE data.

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 Figure 3A) 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 3A, originally computed on the seed sequences only, by analogous histograms computed on all probed natural sequences in the alignment. The resulting histograms are very close to 3A (see Supplementary Figure S2B), and the conclusions of our analysis remain unchanged.
- One can argue that the histograms in Figure 3A might be convoluted by noise, and further improvements could be obtained by attempting to deconvolute this noise. To assess the impact of the noise, we can go in the opposite direction, of further convolving $P(r|\text{bp})$ and $P(r|\text{np})$ by experimental noise and inspecting the effect on the resulting histograms. More precisely, we can resample the reactivities (**12**) from the site distributions $P_{ni}(r|\tilde{r}_{ni})$ and recompute the histograms using these resampled reactivities in place of the original ones. The resulting histograms suffer minor variations in comparison to 3A (see Supplementary Figure S3), and the conclusions of our analysis remain unchanged.
- 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 Figure S2C,D. In both cases, the histograms suffer minor variations, and our overall conclusions are unaffected.

The histograms in Figure 3A do not suffer appreciable changes in all the cases we tested, and our conclusions remain unaffected.

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

$$C^{ab}_{ij}=\frac{1}{B}\sum_n \mathcal{D}^a_{in}D^b_{jn}-\left(\frac{1}{B}\sum_n \mathcal{D}^a_{in}\right)\left(\frac{1}{B}\sum_n \mathcal{D}^b_{jn}\right) \qquad \textbf{(17)}$$

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.

Accompanying software

All the code necessary to reproduce the results in this paper is available as an open-source Julia (2, 9) package. It can be obtained from Github, at the following URL: `https://github.com/cossio/SamApp.jl`.

RESULTS

RBM learns more than secondary structure

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 $\lambda_{\text{reg}}=0.01$. Implementation details of the training can be found in Supplementary Section A. In addition, we have conducted detailed cross-validation analyses supporting these architectural choices, see Supplementary Figure S1.

The riboswitch aptamer structural fold imposes contacts between distant sites along the RNA sequence, which are reflected in concerted covariations between nucleotides in those columns in the MSA. To assess how well the RBM captures sequence features connected to structural constraints, we compute the following epistatic score (73):

$$\mathcal{J}_{ij}=\sum_{a,b}\left\langle \frac{1}{25}\sum_{a',b'}\ln\left[\frac{P(\mathbf{v}^{a,b})P(\mathbf{v}^{a',b'})}{P(\mathbf{v}^{a',b})P(\mathbf{v}^{a,b'})}\right]\right\rangle^2_{\text{MSA}} \qquad \textbf{(18)}$$

for pairs of sites i,j along the sequence. Here, $\mathbf{v}^{a,b}$ denotes a sequence $\mathbf{v}$ from the MSA, which has suffered a double mutation, site i was modified to the letter a, and site j was modified to the letter b. $P(\mathbf{v}^{a,b})$ denotes the likelihood (**5**) 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 (73), is closely related to the Frobenius norm of interactions used in Direct-Coupling analysis for contact prediction in proteins (8), 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 $\mathcal{C}_{ij}$, which has been argued to decrease the impact of phylogenetic biases in contact prediction (8, 16). The APC corrected contact matrix, $\tilde{\mathcal{J}}_{ij}$, is defined by:

$$\tilde{\mathcal{J}}_{ij}=\mathcal{J}_{ij}-\frac{\sum_{kl}\mathcal{J}_{kj}\mathcal{J}_{il}}{\sum_{kl}\mathcal{J}_{kl}} \qquad \textbf{(19)}$$

Figure 4 shows a heatmap of the APC corrected matrix $\tilde{\mathcal{J}}$, for all pairs of positions along the sequence. For some clusters of pairs of sites, $\tilde{\mathcal{J}}$ is significantly larger than for others, suggesting a network of epistatic long-range contacts detected across the sequence. We mapped these predicted contacts to the secondary structure, Figure 4B. It can be seen that large values of $\mathcal{J}_{ij}$ correspond to contacts in the secondary structure. The pseudoknot is correctly detected by this framework, indicating that tertiary contacts are also correctly modeled.

To assess how well the RBM models constraints associated to the secondary structure, we computed the Infernal bit-scores of RBM samples, using the refined CM model. In Figure 4B, we show a scatter plot of Infernal scores versus the RBM effective energies E_{eff} (**6**). RBM sampled sequences have Infernal bit-scores comparable to the natural sequences, indicating that RBM samples satisfy the constraints imposed by the CM model, as well as the natural sequences. On the contrary, Infernal samples have effective RBM energies significantly higher than the natural sequences (orange threshold in the panel), suggesting that the RBM imposes further constraints beyond those imposed by the CM, such as tertiary motifs (pseudoknot, A-minor motif), and possibly other functional constraints not representable by the CM, and which are however important for the function of the aptamer. To evaluate the impact of the pseudoknot on this energetic gap, we repeated this experiment using the untangled CM instead of the refined CM. Results are collected in Supplementary Figure S9. The untangled CM samples sequences with better complementarity and Turner energies favorable for base-pairing along the pseudoknot. However, an RBM energetic gap of similar magnitude as in 4B is still observed, suggesting that additional tertiary motifs or functional constraints might be at play. Similar results are also obtained with the Rfam CM, see Supplementary Figure S9.

We then inspected the hidden units of the model. We find that hidden units learn extended features, corresponding to long-range interactions along the sequence. Hidden units #44 and #67 are found to have weights with the highest norms, and are representative of the other hidden units of the

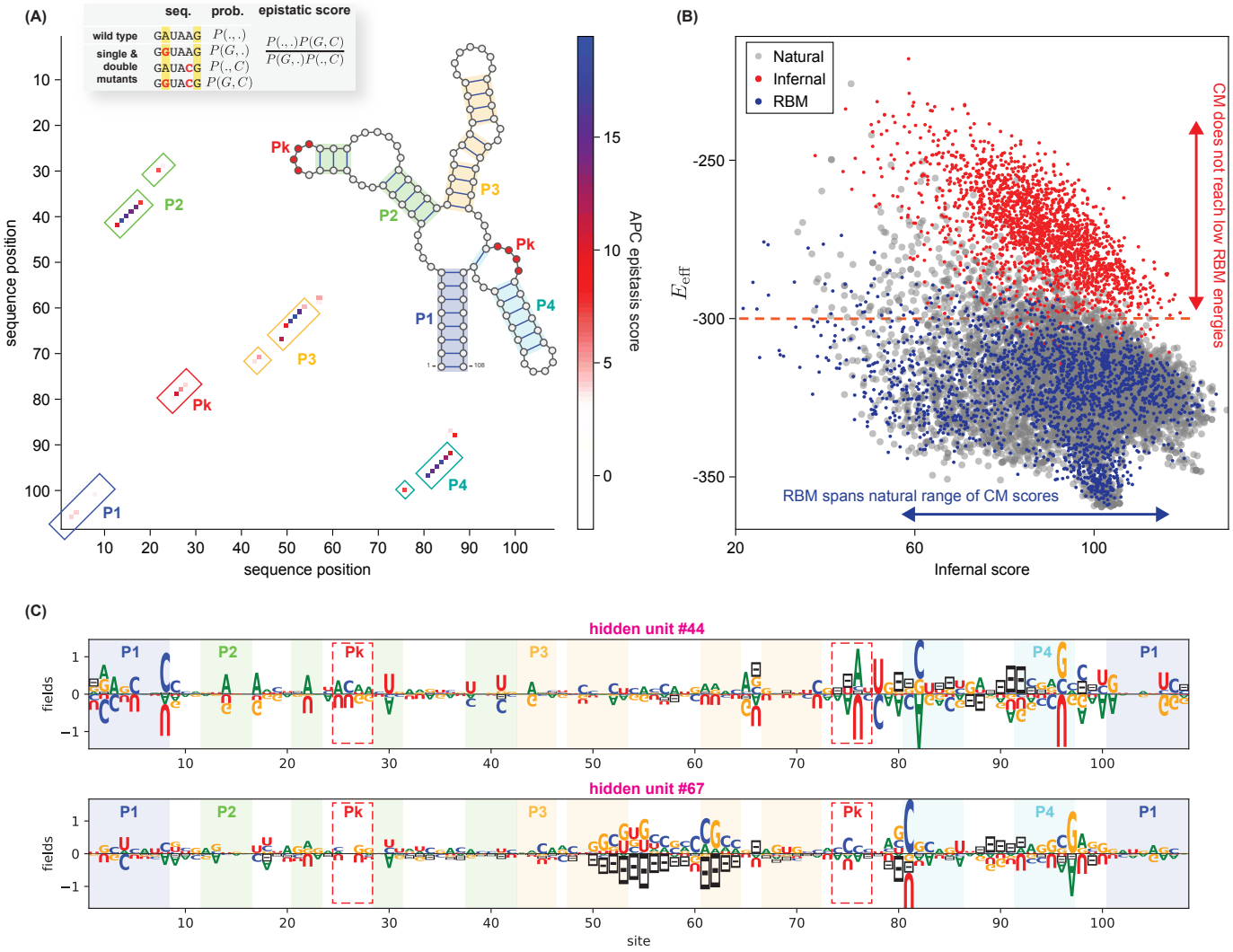

Figure 4. How the hidden units encode sequence constraints relevant for function. **(A)** Contact map, constructed by computing an epistatic score from the RBM marginal probability over sequences, see inset table and Equations (**18**) and (**19**). The highest epistatic scores correspond to secondary and tertiary contacts of the SAM-bound aptamer structure, shown in the inset. **(B)** Scatter plot of Infernal bit-scores (x-axis) vs. RBM effective energies E_{eff} (y-axis), for natural sequences (gray) alignment, Infernal sampled sequences (from the refined CM model, in red), and RBM sampled sequences (blue). A threshold at $E_{\mathrm{eff}} = -300$ (orange dashed line) highlights a separation in the effective energies assigned by the RBM to the Infernal generated sequences, in comparison to natural and RBM samples. **(C)** Logos of the weights $w_{i\mu}(v_i)$ attached to exemplary hidden units ($\mu = 44$ and $\mu = 67$), selected by having the highest weight norms). Sites are annotated by the secondary structure element to which they belong with different colors, including the paired (P) helices P1 (light purple), P2 (green), P3 (yellow), and P4 (teal). The sites participating in the pseudoknot (Pk) are also highlighted (red dashed box). The hidden units capture extended motifs, including long-range contacts relevant for secondary and tertiary structure formation (as discussed in the text).

model. Figure 4C plots the weights attached to these units, represented as sequence logos. In the plot, the size of a letter is proportional to the corresponding weight, with letters below zero corresponding to negative weights (see (69) where a similar representation is used). In hidden unit #44, Watson-Crick complementarity between sites 9 and 101 is favored, which is compatible with the base-pairing of these positions at the 5' and 3' ends of the P1 helix. The same unit also places a large weight on complementarity between sites 25-26 and 76-77, helping stabilize the pseudoknot tertiary contact. The fact that these complementarity constraints, belonging to different structural motifs, are enforced by the same unit, is compatible with the notion that P1 and the pseudoknot are two structural elements that stabilize in closed conformation in a concerted manner, in response to SAM. Hidden unit #67, on the other hand, places significant weight in the complementarity between sites 81 and 97, stabilizing P4. This hidden unit also places moderate weights distributed along various positions of P3, favouring a dichotomy between stabilizing complementarity or deletions in this segment. Indeed, many of the natural sequences lack the hairpin loop of P3 (sites 50–64), in a manner compatible with the snegative activation of hidden unit #67.

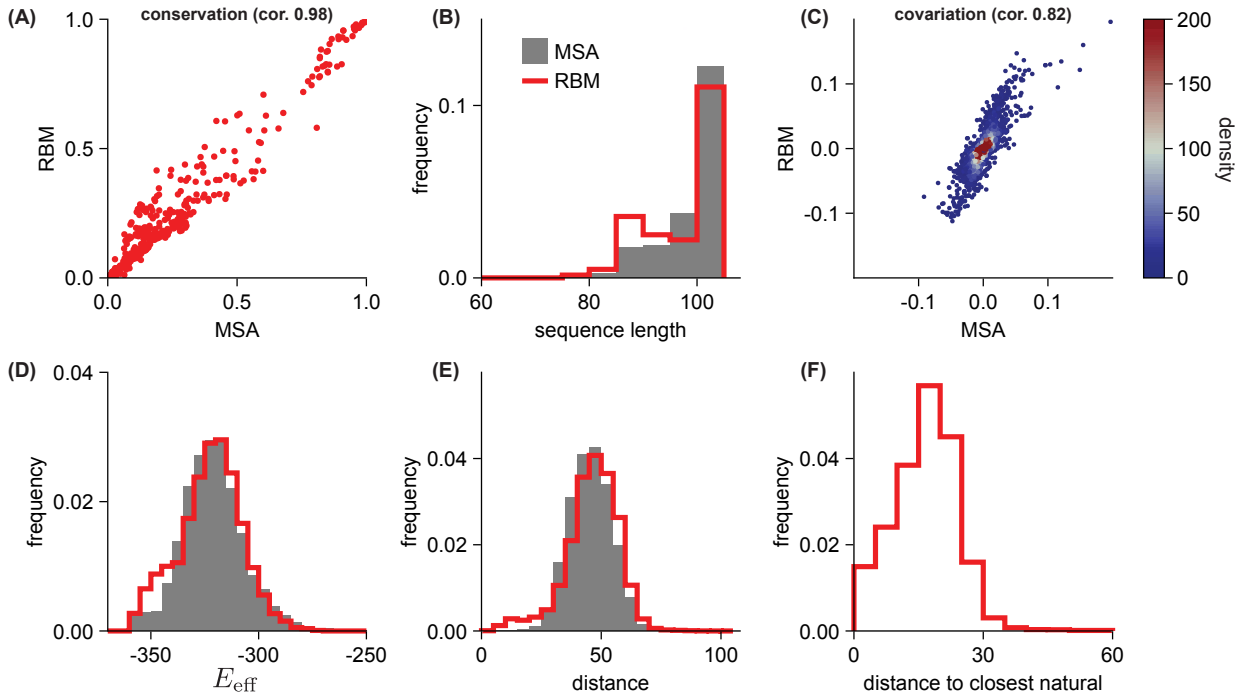

Figure 5. RBM generates novel and diverse sequences that recapitulate statistics of natural homologues. **(A)** RBM samples reproduce single-site nucleotidic conservation of the natural sequences (Pearson correlation $=0.98$). **(B)** Histogram of natural sequence lengths (gray) and of RBM generated sequences (red). Note that insertions are discarded. Sequence length is defined as the number of aligned sites that are not gaps (deletions). **(C)** RBM samples reproduce the statistics of nucleotidic covariation of natural sequences (Pearson correlation $=0.82$). Since the number of paired sites is very large, the points are colored by their density in the plot, according to the colorbar legend. **(D)** Histograms of effective energies E_{eff} (**6**) of natural sequences (gray) and of RBM samples (red). **(E)** Histograms of pairwise Hamming distances, among natural sequences (gray) and among RBM samples (red). **(F)** Histogram of Hamming distances, from each RBM sampled sequence, to its closest natural sequence.

RBM generates diverse sequences compatible with the statistics of natural homologs

The quality of the model fit after training, and the quality of convergence, can be assessed by comparing statistics of sampled sequences against the empirical statistics of the MSA. In Figure 5A, we compare the single-site statistics, computed as the frequency of occurrences of each nucleotide (or gap symbol) at each position of the alignment. The agreement is excellent (Pearson correlation $=0.98$), indicating that the RBM reproduces the conservation of important sites (cf. Figure 1C). Furthermore, RBM sampled sequences reproduce the covariance of pairs of sites of the natural sequences, as shown in Figure 5B. In this case, we compute the deviation of the frequencies of co-occurring nucleotides at pairs of sites from the expectation arising from their independent conservations. Such joint covariations arise from interactions across the sequence, related for example to secondary or tertiary contacts, or other functional constraints. The agreement is also excellent (Pearson correlation $=0.97$), indicating that the RBM is able to reproduce the covariation of the natural MSA. We also evaluated the RBM effective energies E_{eff} (**6**) of sampled sequences compared to the energies assigned by the RBM to the natural sequences. As we show in Figure 5C, the two histograms are in close agreement to each other.

To evaluate the diversity of a set of sequences, natural or generated, we compute the matrix of all possible pairwise Hamming distances between pairs of distinct sequences, where the Hamming distance is defined as the number of positions where the two sequences differ. Figure 5D shows the histogram of these pairwise distances for the natural sequences in gray. Typically, two randomly selected natural sequences differ in about 40% of sites, or 43 out of the 108 aligned sites. We then sampled 5000 sequences from the RBM, and computed the histogram of their pairwise distances (between themselves). We plot the result in Figure 5D in red. We see that the histogram closely resembles the histogram of the natural sequences. We conclude that the RBM generated sequences recapitulate the natural diversity of the sequence homologs family. Furthermore, the RBM generates novel sequences, not seen in the data. Indeed, Figure 5E shows the histogram of distances between each sampled sequence, and the closest natural sequence to it. Typical RBM samples differ in 20 sites from the closest sequence in the MSA, and therefore constitutes a truly novel sequence.

Finally, we observed a strong variation in sequence lengths in natural sequences. In particular, dramatic variations of the P4 helix have been reported in the literature (26, 72), where riboswitches without P4 have been shown to be functional although with lower affinities to SAM. Although our RBM is not able to model insertions, it is still able to emit sequences of varying lengths by having more or less gaps in the sequence. We therefore compared the distribution of sequence lengths generated by the RBM, with the histogram of natural sequence lengths (not considering inserts) from the MSA. The plot

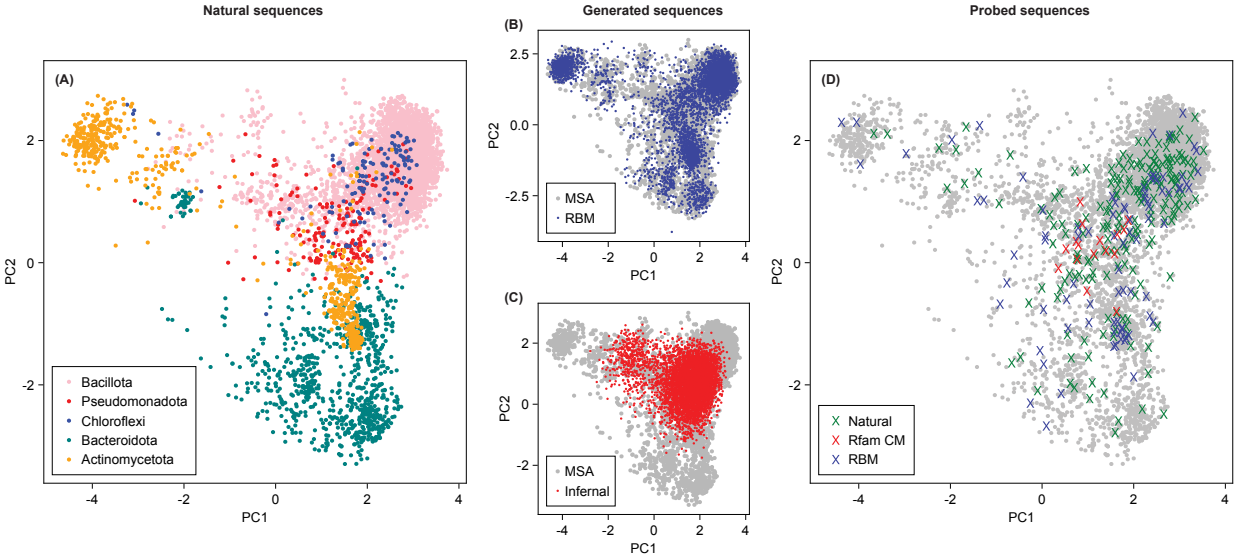

Figure 6. Principal component analysis of natural MSA and generated sequences. **(A)** Projection of natural sequences of the full MSA onto the top two principal components of the correlation matrix of the MSA. The largest taxonomic groups (with >100 member sequences) are shown. Taxonomic annotations were fetched from NCBI. **(B)** Projection of RBM sampled sequences on the top two principal components of the MSA (in red), with the natural sequences shown in background (gray). **(C)** Projection of Infernal sampled sequences on the top two principal components of the MSA (in red), with the natural sequences shown in background (gray). **(D)** Projections of the sequences probed in our experiments, divided according to their origin: Natural (green), sampled from Rfam CM (red), and sampled from the RBM (blue).

in Figure 5F confirms that the RBM reproduces the correct length statistics.

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

To further evaluate the sequence space coverage of RBM samples, we considered a principal component analysis (PCA) of the full MSA sequences. The top two principal components (PC) are shown in Supplementary Figure S5. The top principal component captures a mode of variation associated to deletion of the helix P4, as can be appreciated from the large number of gaps in this region. We project every sequence of the MSA onto the top two principal components. We obtained the taxonomic class annotations of all sequences in the MSA from Rfam. Figure 6A shows the projections of all natural sequences onto the top two components, with colors corresponding to the most populated taxonomic classes. The principal components appreciably separate taxonomic clusters of natural sequences. In particular, a group of Actinomycetota aptamers, in the top left corner of Figure 6A, have completely deleted or very short P4 helix segments. Literature reports suggest that SAM aptamers are able to function in absence of P4 (72), although the affinity towards SAM decays with decreasing P4 length (26). We then projected RBM sampled sequences onto the two principal components, see Figure 6B. The RBM samples span the full space, covering all the clusters associated to different taxonomic classes. In contrast, Infernal-generated sequences, shown in Figure 6C, span a limited region of the principal component space, and remain confined to a central location of the PCA plot. Since Infernal is able to model a limited amount of covariation, this suggests that the flexibility of RBM to capture further underlying interactions in the data is important to model the full span of the family of homologues. Finally, Figure 6D shows the projection of the sequences probed in our SHAPE-MaP experiments.

Selection of sequences for experimental probing

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 effective energy of the sample. We found that 5000 steps were more than sufficient. We then sorted these sequences by their value of E_{eff}, and selected 70 sequences at random, uniformly spanning the range of‘ energies observed in the sample. The table of sequences and their associated effective energies is reported in the Supplementary Materials.

We hypothesized that functionality of RBM sequences should strongly correlate with their effective energy E_{eff}. We therefore define a group having *low*-RBM effective energies, with $E_{\text{eff}} < -300$ and consisting of 53 sequences. We also define a group of *high*-RBM effective energy sequences, having $E_{\text{eff}} > -300$, and containing 31 sequences. Setting energetic thresholds is similar to sampling the model at a temperature different from 1, also found to be helpful in DCA models used for design (56). The chosen threshold value $E_{\text{eff}} = -300$ approximately separates the bulk of natural and RBM sampled sequences, from the CM samples (Figure 4B).

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 55 sequence members of the hits MSA, and 151 sequence members of the seed MSA (not part of the seed), as described in Methods. The selected natural sequences are diverse, spanning various taxonomic classes (see Figure 6D). A listing of probed sequences can be found in the Supplementary Material.

Average SHAPE reactivity response to SAM

Reactivity measurements can be subject to significant noise, related to sequencing errors and other biases (62). We can counteract the impact of noise by averaging multiple measurements. We have here computed the average reactivity over groups of related sequences, for each site. For seed MSA sequences,

$$\langle r_i \rangle_{\text{seed}} = \frac{1}{B_{\text{seed}}} \sum_{n \in \text{seed}} r_{ni} \qquad (20)$$

where B_{seed} is the number of sequences in the seed alignment. Similarly, we computed averages over the RBM designed sequences with low effective energy ($E_{\text{eff}} < -300$), RBM designed sequences with high effective energies ($E_{\text{eff}} > -300$), and sequences sampled from the Rfam CM model with Infernal. Then, to evaluate the response of the aptamers to SAM, we computed for each aptamer and each site, the difference in reactivity in the condition Mg+SAM minus the condition with Mg only.

Figure 7A shows the average reactivity difference of natural sequences from the seed alignment (in gray). The full MSA has a similar behavior (see Supplementary Figure S7). The response to SAM is appreciable from the significant protection of sites related to SAM binding and the structural switch. Here we observe a reactivity decrease at the sites near the pseudoknot (25-28, 77-80, labeled in red), and sites directly in contact with SAM or forming the A-minor tertiary structural motif (as indicated by the green arrows). This indicates that, broadly sepeaking, sequences from the seed alignment are on average SAM binders, as expected. The figure also shows the bands indicating the standard deviation of the reactivity in the group of sequences. Although these deviations are small, the average behavior shown here does not preclude the possibility that some sequences might fail to respond to SAM.

Figure 7A overlays in blue the average reactivity difference of RBM sequences with low effective energy ($E_{\text{eff}} < -300$). The average delta-reactivity profile of these RBM generated sequences is in excellent agreement with the profile of the natural sequences, exhibiting the protection that implies a significant response to SAM. This shows that RBM designed sequences of low enough energy are able to recapitulate the behavior of their natural counterparts.

Figure 7B then shows the delta reactivities of RBM sequences with high energies ($E_{\text{eff}} > -300$) in red, overlaid on top of the natural profile (gray). This group of sequences shows appreciable lack of protection at key sites, such as 10-11 (SAM contact), 73-76 (Pseudoknot and A-minor), and 97-101 (SAM contact at P1). These sequences therefore show no or weak response to SAM, indicating poor binding and lack of a structural switch necessary for regulatory function. A similar behavior is exhibited by Rfam CM sampled sequences with Infernal, shown in Figure 7C. Again, these sequences seem to be non-functional.

We can also confirm that the RBM sequences with $E_{\text{eff}} < -300$ reproduce the reactivity response to magnesium of natural sequences. See Supplementary Figure S8. In contrast, RBM sequences with higher energies ($E_{\text{eff}} > -300$) show larger discrepancies. We conclude that low RBM energy aptamers recapitulate structural responses to both SAM and magnesium, of natural sequences.

SHAPE reactivies are in broad 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 tengcongensis* and the *Bacillus sub- tilis yitJ* SAM riboswitches, which have been extensively studied in the literature fueled by detailed knowledge of their published crystalized structures (36, 43). For many 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 posited for the RF00162 family, shown in Figure 1B. Indeed, we have computed the average evidence scores $\langle (i)S \rangle$ for each site i, over the sequences in the seed alignment probed in our experiments,

$$\langle \mathcal{S}(i) \rangle_{\text{seed}} = \frac{1}{B_{\text{seed}}} \sum_{n \in \text{seed}} \mathcal{S}_n(i) \qquad (21)$$

Figure 8 plots $\langle \mathcal{S}(i) \rangle$ over the different conditions in our experiment: no SAM and no Mg (30C), in presence of magnesium (Mg) only, and with both magnesium and SAM (SAM+Mg). Overall, the averaged scores are in detailed agreement with the consensus secondary structure of the aptamer, depicted in Figure 8A. 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 (43)), and other tertiary motifs known to form in response to SAM. We discuss these next.

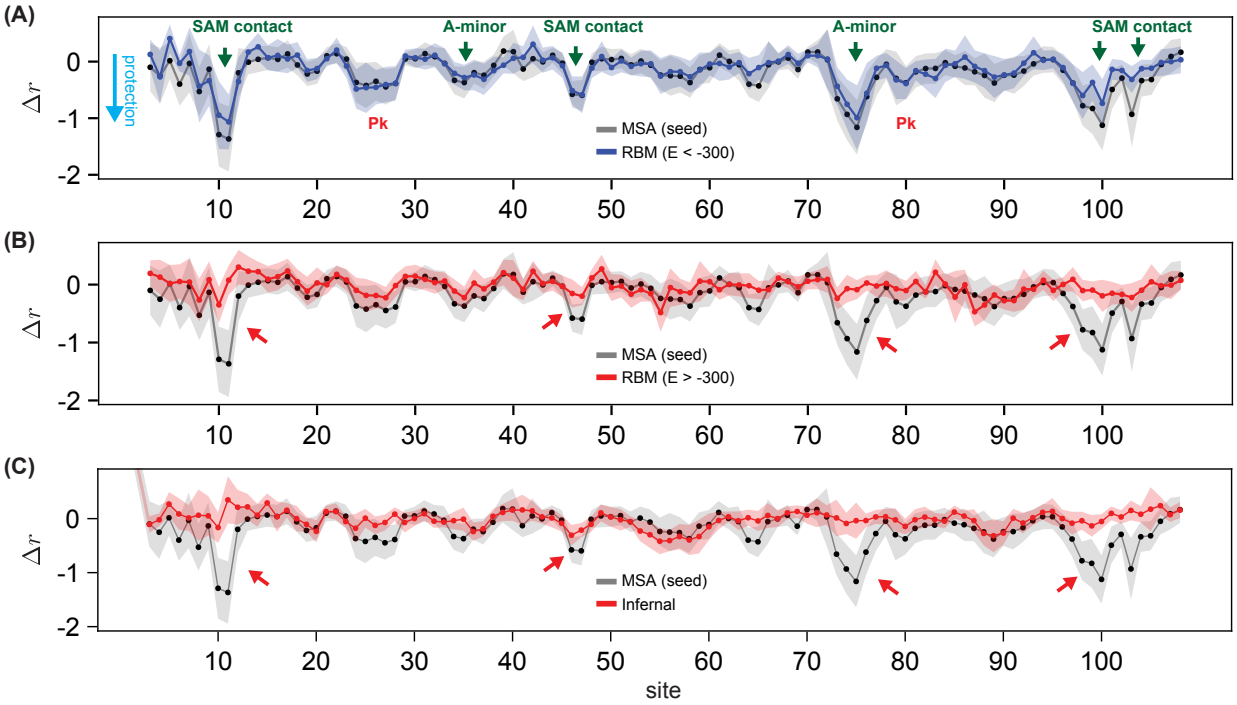

Figure 7. Average differential reactivities in response to SAM: **(A)** for natural sequences (in gray) and low-E_{eff} RBM generated sequences (in blue; $E_{\text{eff}} < -300$), **(B)** for natural sequences (in gray) and low-E_{eff} RBM generated sequences (in red; $E_{\text{eff}} > -300$), and **(B)** for Infernal generated sequences, using the Rfam CM. For each group of sequences and each site, we computed the average reactivity difference of the condition with SAM+Mg minus the condition with Mg only. The plots show the average profiles of sequences in the group, with the bands indicating $\pm$ half one standard deviation. The green arrows indicate locations expected to exhibit reactivity differences in response to SAM binding, while the red arrows highlight lack of protection in some groups of sequences in those locations.

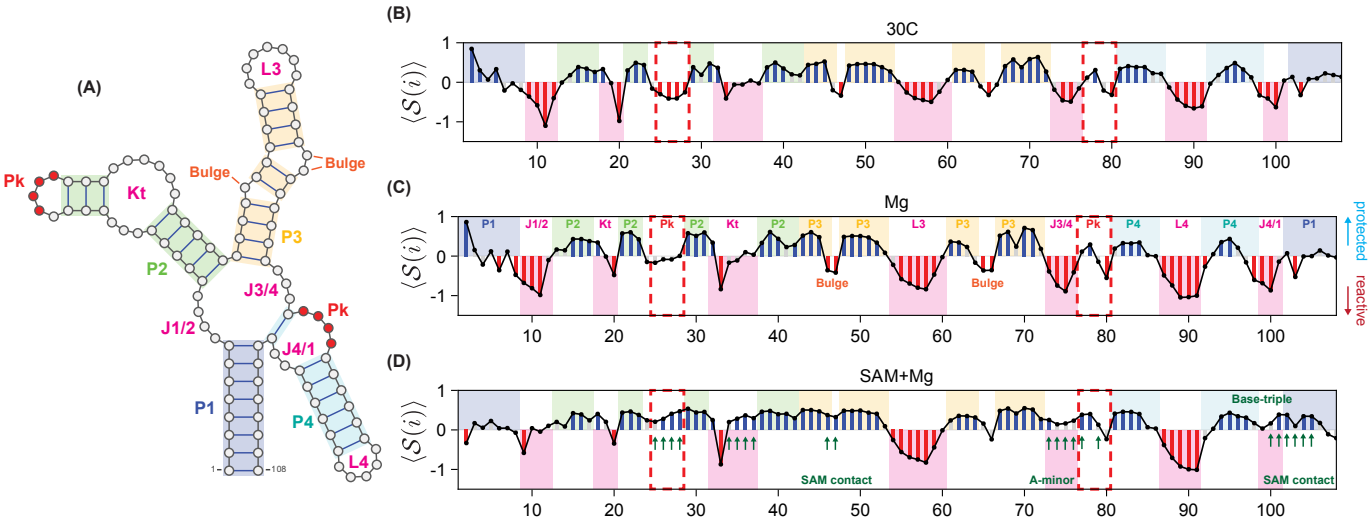

Figure 8. Average reactivity scores, $\mathcal{S}(i)$ (see (**16**)), per site of probed sequences, for the three conditions: no SAM and no Mg (30C), in presence of magnesium (Mg), and in presence of magnesium and SAM (SAM+Mg). **(A)** Annotated secondary structure. The colors are the same used in the other panels. **(B)** Average $\mathcal{S}(i)$ over probed sequences without SAM and without Mg (30C). **(C)** Average $\mathcal{S}(i)$ over probed sequences, in presence of magnesium (Mg). **(C)** Average $\mathcal{S}(i)$ over probed sequences, in condition with SAM and magnesium (SAM+Mg). In B, C, and D, the annotations indicate locations of important structural elements, with colors matching panel A: helices P1, P2, P3, P4, four-way junction (J1/2, J3/4), loops L3, L4, the pseudoknot (Pk), a kink-turn (Kt), the junction between P4 and P1 (J4/1), and bulge sites (orange). In addition to the pseudoknot (Pk, in red), other sites are expected to suffer changes in reactivity in response to SAM binding, forming additional tertiary contacts, such as a base-triple and the A-minor triple. These are indicated by green arrows in D, and discussed in the text. See also Table 1.

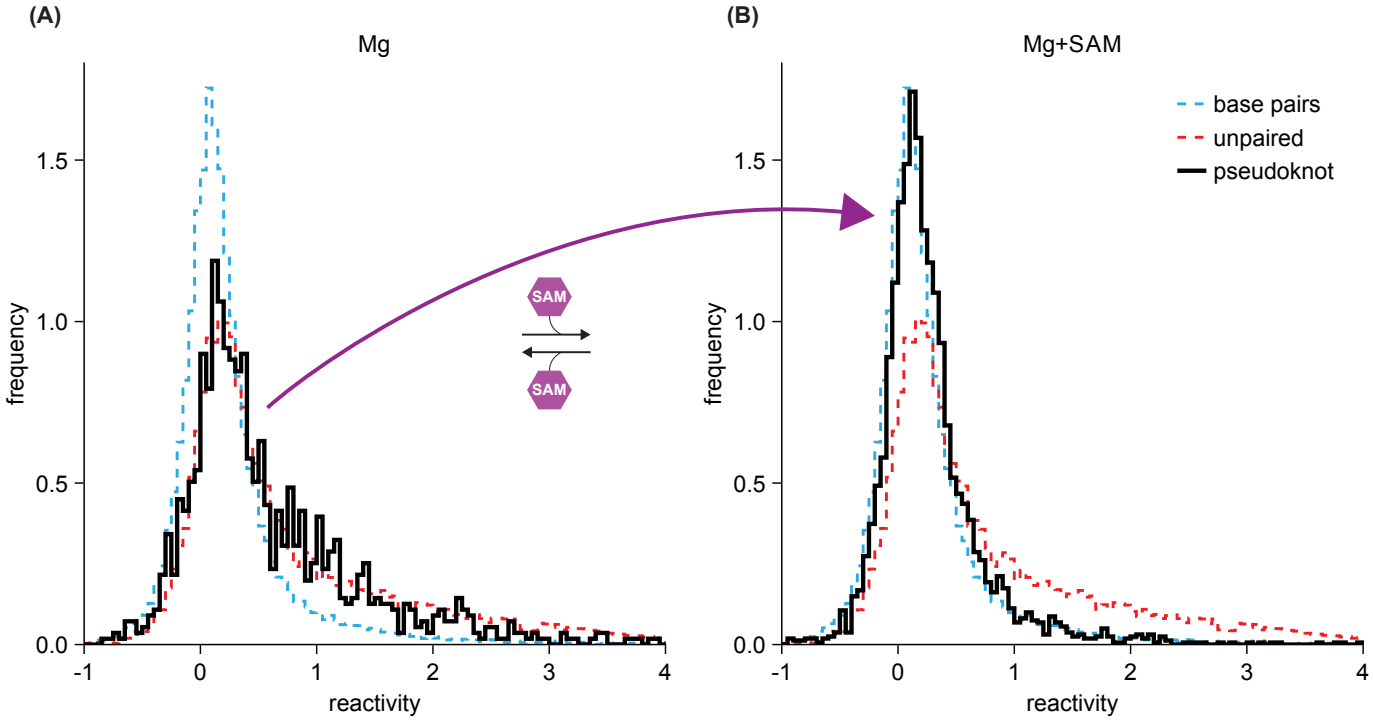

Figure 9. Pseudoknot formation is reflected in SHAPE reactivities. **(A)** Histogram of reactivities of base-paired (blue) and unpaired sites (red) in probed natural sequences belonging to the manually curated seed alignment (cf. Figure 3). Histogram of pseudoknot sites, in absence of SAM, is shown in black, and agrees with the distribution of

SHAPE reactivities broadly support formation of pseudoknot in natural sequences upon SAM binding

In addition to being compatible with expected secondary and tertiary structural motifs, natural sequences are broadly able to respond to SAM by performing a structural switch. In particular, it is recognized in the literature that SAM riboswitches stabilize a pseudoknot in response to SAM which helps create a binding cradle for the ligand, as well as other tertiary contacts. Having established reactivity histograms associated to base-paired and unpaired sites (see Figure 3), we reasoned that pseudoknot sites should have, in absence of SAM, reactivities compatible with unpaired sites, while in presence of SAM, the pseudoknot reactivities should instead approach reactivities of base-paired sites. The plots in Figure 9 confirm these expectations. We find that in absence of SAM (left panel), reactivities of pseudoknot sites are in excellent agreement with the unpaired reactivity histogram (red), while in presence of SAM (right panel), the pseudoknot reactivities shift towards the base-paired histogram (blue). We find a similar result for the P1 helix, see Supplementary Figure S4.

This result supports two conclusions: that, in spite of possibly including non-functional sequences or other biases, the histograms in Figure 3A are accurate enough approximations of the underlying probability distributions of paired and unpaired reactivities; and second, that indeed most natural sequences in the RF00162 alignment (seed and full) probed in our experiments bind SAM and respond by the expected structural switch. We remark that this statement is only valid as an average over the bulk of natural sequences, and does not exclude the possibility that some particular sequences fail to bind SAM, or respond in a different manner.

Table 1. Hallmark sites of structural switch in response to SAM

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

List of sites that exhibit observable SHAPE reactivity differences upon SAM binding. Positions are numbered following the Rfam reference alignment. See also Figure 8.

Structural switch in response to SAM is reflected in reactivity changes of hallmark sites

Based on these observations, we selected a number of hallmark sites across the aptamer sequence, that have observable reactivity changes in response to SAM binding, and are consistent with expectations from previous published studies on SAM-I riboswitches. These sites are listed in 1, 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 (39), to crystal structure (36, 43), to

sequence based statistical modelling (79). Consistent 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 Figures 7 and 8D. We therefore include these sites in Table 1.

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, (43)) 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+}.

A-minor. The A-minor motif helps create a groove where SAM is placed upon binding. It is an important ligand-dependent structural element (36, 43, 65). 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.

Base-triple. The base-triple is an important tertiary motif observed in the bound structure of the aptamer (36, 43), 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 (25, 43). 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 (65). Stabilization of this tertiary structure in response to SAM has been observed, and evidenced both in the crystal structure (36, 43, 59), previous SHAPE experiments (25), and simulations (65). 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 1.

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 (43). 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 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 Figure 1A), we expect significant reactivity decreases in this region as well upon SAM binding.

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 Figure 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 (43).

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.

Low effective RBM energy correlates with structural response to SAM

Having established a set of hallmark sites that exhibit robust reactivity responses to SAM in most natural sequences, see Table 1, we proceeded to exploit the reactivity measurements at these sites to classify the behavior of all the probed sequences in response to SAM. We computed a motif evidence score $\mathcal{S}(\mathcal{M})$, as defined in (**16**), where $\mathcal{M}$ is the set of hallmark sites identified in Table 1. Figure 10 plots $\mathcal{S}(\mathcal{M})$ in the Mg and SAM+Mg conditions, against the RBM effective energy $E_{\rm eff}$ of the sequences probed in our experiments. The structural switch in response to SAM, is manifested in these hallmark sites, by a shift in $\mathcal{S}(\mathcal{M})$ from negative values in presence of Mg (indicating these sites are likely to be unpaired), to positive values in presence of both SAM+Mg (indicating these sites are now likely to be paired). We consider a threshold of $\ln(5)$ for a 5-fold statistical significance (shown as the yellow dashed lines in the figure). Sequences that respond to SAM are shown as filled disks in the figure. A summary of our results is presented in the Table 2.

We find that SAM responding sequences (both natural and designed) tend to have low RBM effective energies. In particular, 47% of RBM designed sequences below an energy threshold of $E_{\rm eff} < -300$ are functional aptamers, exhibiting significant response in the hallmark sites. These sequences have Hamming distances between 10 and 20 to the closest natural sequence, see Supplementary Figure S10. There are two modes of failure for the remaining RBM sequences: either the structural motifs (pseudoknot, P1, etc.) are always paired regardless of whether SAM is present or not, or the structural motifs never close. We find that the 55 RBM sequences that we conclude are non-responsive to SAM fail to do so in the second manner: they do not form the necessary contacts with or without SAM. Sequences sampled from the Rfam CM are non-functional, possibly on account of the inability of the CM to model tertiary contacts like pseudoknots (17, 46).

Stabilization of structural elements in response to SAM is consistent with energetic calculations

SAM binding results in a dramatic reorganization of the structure of the aptamer. Formation of a pseudoknot and stabilization of the P1 helix are two hallmarks of this transition. We reasoned that, in order to respond to SAM, the molecule must have a certain degree of structural flexibility

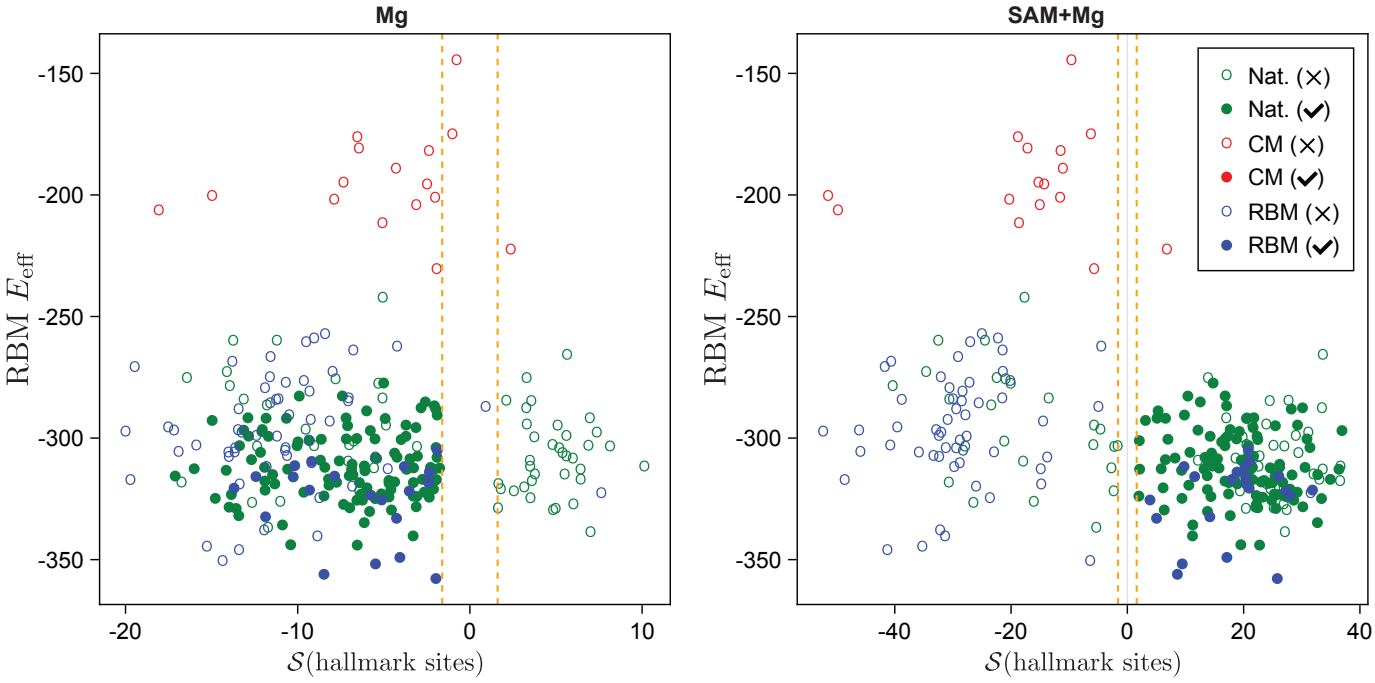

Figure 10. Sequences with low RBM E_{eff} perform structural switch in response to SAM. Both panels show scatter plots of the log-odds score $\mathcal{S}$ (**16**) evaluated on the switch hallmark sites identified in Table 1 on the x-axis, against the RBM effective energy E_{eff} (**6**) on the y-axis. In the left panel, $\mathcal{S}$ is evaluated with the SHAPE data of probed sequences in the condition with magnesium (Mg), while the right panel considers probing in presence of SAM and magnesium (SAM+Mg). The filled points correspond to functional sequences for which we observe a significant response to SAM: $\mathcal{S} < -\ln(5)$ in the Mg condition, and $\mathcal{S} > \ln(5)$ (five times more likely to be unpaired) in the condition with SAM+Mg (five times more likely to be paired). In other words, these are the sequences that responde to SAM with significant reactivity changes in the selected sites. Sequences are colored according to their origin: Natural sequences (seed + full) are in green, Rfam CM generated sequences in red, and RBM sequences in blue. The pink dashed line marks the energetic threshold $E_{\text{eff}} = -300$.

Table 2. Experimental functionality of different groups of probed sequences

Sequence group	Number (Conclusive)	Responsive to SAM (% of conclusive)	Not responsive to SAM (% of conclusive)
Natural	206 (170)	109 (64%)	61 (36%)
Seed MSA	151 (134)	90 (67.1%)	44 (32.8%)
Full MSA	55 (36)	19 (52.8%)	17 (47.2%)
Natural ($E_{\text{eff}} < -300$)	137 (121)	84 (69.4%)	37 (30.6%)
Natural ($E_{\text{eff}} < -310$)	96 (86)	63 (73.3%)	23 (26.7%)
RBM	84 (76)	21 (28%)	55 (72%)
RBM ($E_{\text{eff}} < -300$)	53 (45)	21 (46.7%)	24 (53.3%)
RBM ($E_{\text{eff}} < -310$)	40 (32)	19 (59.4%)	13 (40.6%)
Rfam CM	16 (16)	0 (0%)	16 (100%)
All	306 (262)	130 (49.6%)	132 (50.4%)
All ($E_{\text{eff}} < -300$)	190 (166)	105 (63.3%)	61 (36.8%)
All ($E_{\text{eff}} < -310$)	136 (118)	82 (69.5%)	36 (30.5%)

For the different groups of sequences (rows), the columns show: the total number of sequences probed in that group (with the number for which the experimental measurement was conclusive in parenthesis), the number of sequences which were responsive to SAM by significant reactivity changes in the Hallmark sites identified in Table 1, and the number that were identified as not responsive.

to be compatible with the two competing folds and be able to switch from one to the other. We computed the binding energies of our probed sequences across the P1 helix and the pseudoknot sites, using the Turner energy model as implemented in the ViennaRNA package (35). We observe that some sequences respond in one or more of these structural elements when SAM is bound, by moving from a conformation where the helix is unstable to a state where the helix is stabilized. In order to implement this switch, molecules cannot have a binding energy too low (or the motif can never close), nor too high (or the motif will always be closed). Figure 11 shows that the sequences that respond to SAM across these motifs are confined to a binding energy window from -10 to 0 kcal/mol for P1, and between -8 and -3 kcal/mol for the pseudoknot (Pk). Since P1 consists of 16 base-paired site, while the pseudoknot involves 8 sites, in both cases the flexible energy band spans a range of 1.25 kcal/mol per base-pair, or about one third of the typical base-pair energy (35).

Sequences with energies above this window, tend to have the structural element always open, while sequences with energies below this window, have the structural element

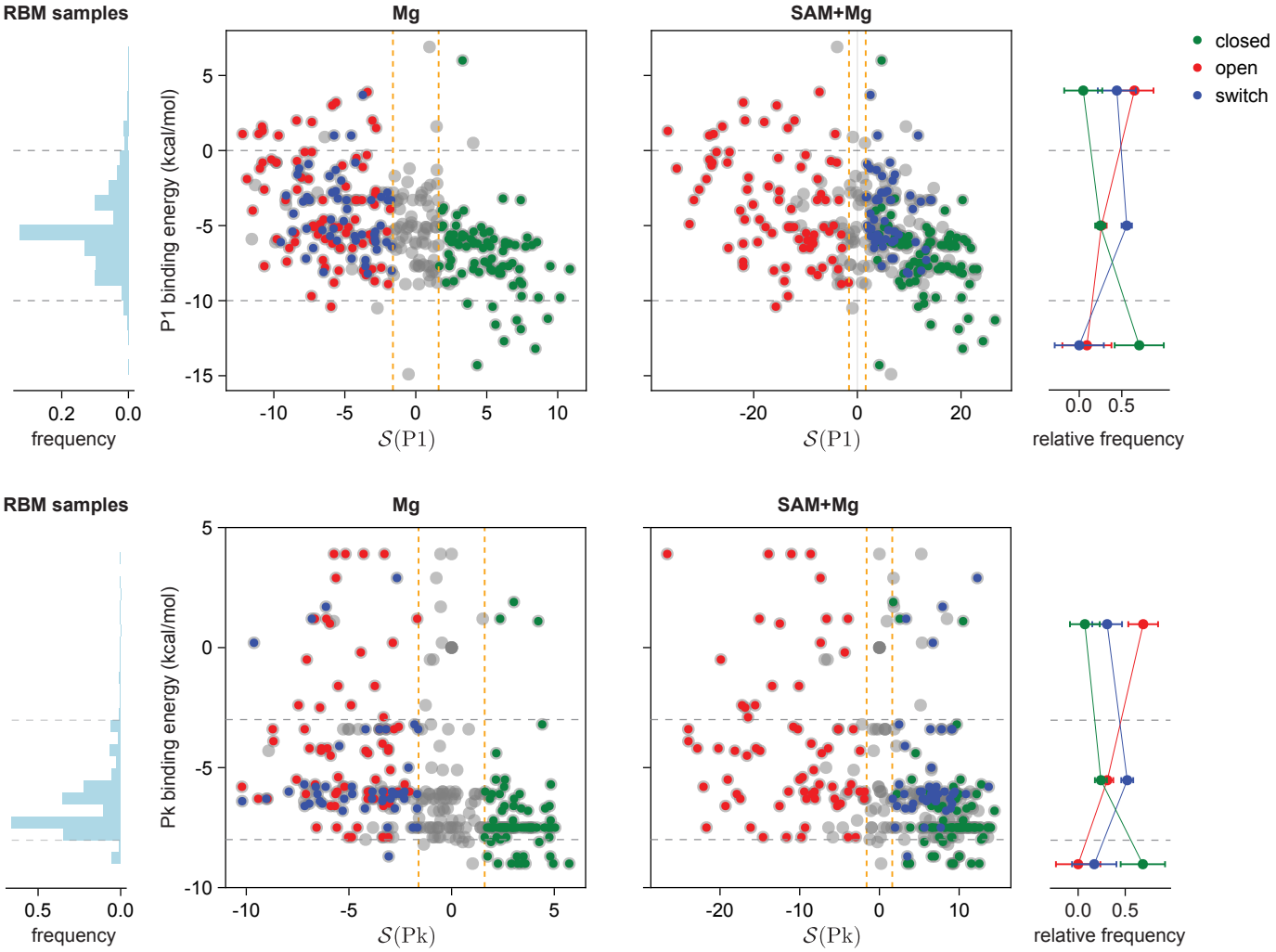

Figure 11. SAM stabilizes structural elements of the aptamer with intermediate binding energies. We consider the P1 helix (top row) and the pseudoknot (Pk, bottom row). Leftmost panels show the histogram of Turner binding energies (computed with the ViennaRNA package (35)) associated to the motifs (P1 or pseudoknot), of sequences designed by the RBM. The second and column plots on the x-axis the base-pairing log-odds ratio of the motifs, $\mathcal{S}$(P1) for the P1 helix on the top and $\mathcal{S}$(Pk) for the pseudoknot on the bottom (defined in Equation (**16**)), versus the Turner binding energy. In the second column, $\mathcal{S}$ is computed in the condition with magnesium, while in the third column $\mathcal{S}$ is computed in the condition with SAM and magnesium (SAM+Mg). The aptamers are colored according to their behavior in response to SAM: i) if $\mathcal{S} > \ln(5)$ in both conditions, the motif (P1 or Pk) is always closed (green); ii) if $\mathcal{S} < -\ln(5)$ in both conditions, the motif is always open (red); iii) if $\mathcal{S} < -\ln(5)$ with Mg but $\mathcal{S} > \ln(5)$ with SAM, the motif responds to SAM by closing (blue). Finally, points for which we cannot establish their structural states with over 5-fold confidence are shown in light gray. To evaluate how the binding energy impinges on the aptamer response, the fourth column shows the frequency (relative to the total number of aptamers) of each response type in three energetic bands. Error bars are computed based on the number of probed aptamers in the corresponding energetic band. Consistent with biophysical expectations, we find that always closed aptamers tend to have low binding energies, always open aptamers have high binding energies, and aptamers that respond have intermediate binding energies.

always open. Only sequences within this energetic window are flexible enough to transition from one conformation to the other. An optimal binding energetic band is also observed in enzyme function, known as Sabatier' principle.

We then considered a uniform sampling of sequences generated by the RBM. The leftmost panel of Figure 11 shows that RBM samples preferentially have binding energies in the expected intermediate band, and are thus compatible with the structural switch required for riboswitch function.

In Figure 11, we find that 47 probed sequences close the pseudoknot in response to SAM, and 54 probed sequences respond to SAM by closing P1. Interestingly, 46 out of the 47 sequences that pair Pk when SAM binds, are also functional in the sense of Figure 10 and show responses in the sites in Table 1. In our data we don't observe any sequence that closes the pseudoknot but is found to be non-functional. This is consistent with the known importance of the pseudoknot in the response of the aptamer. P1 is somewhat more ambiguous. We find 6 sequences that respond to SAM by stabilizing P1 in Figure 11, but they are classified as non-functional in Figure 10, indicating that these 6 sequences do not display significant reactivity changes in the other sites of Table 1. Therefore the stabilization induced by SAM in these sequences is incomplete. These results are summarized in Table 3.

Table 3. Aptamers response to SAM in structural elements

Hallmark resp.	Total	P1 switch	P1 rigid	Pk switch	Pk rigid
Yes	126	40	47	46	27
No	135	6	98	0	103

Comparison of structural response to SAM in hallmark sites, and response in the structural elements P1 and the pseudoknot. The first row corresponds to the sequences that respond to SAM in the hallmark sites defined in Table 1, while the second row corresponds to sequences that do not exhibit response at these sites. The following columns then show how many of these sequences also exhibit a switching response in P1 or the pseudoknot (Pk). Note that the sums of the later columns (P1 switch + P1 rigid, or Pk switch + Pk rigid) gives the number of sequences for which we are confident in the switching response of the motif (P1 or Pk) and the hallmark sites, and is therefore less than the Total in the second column.

DISCUSSION

The design of small regulatory RNAs has many applications in developing laboratory tools for gene function studies and in drug design, as they can be used to regulate gene expression. Being able to design allosteric and regulatory RNA is also at the core of DNA-RNA computing, and of the investigation of possible scenarios for the origin of life (6, 28, 60).

In this work, we have studied the homologue sequence family of aptamer domains of SAM-I riboswitches. We have shown that RBM models, learned from sequence data, are effective as generative models, able to design artificial SAM-I riboswitch aptamers that successfully transition between conformations upon SAM binding.

To probe artificial sequences we have carried out a high throughput SHAPE-MaP (62) screening of many sequences, including both natural belonging to the SAM-I riboswitch aptamer family, and artificial generated by the RBM and a CM model. We have developed a statistical pipeline to analyze the measured SHAPE reactivities, which takes advantage of the closely related statistics of the ensemble of tested sequences and their shared consensus secondary structure. Our analysis does not rely on a biophysical implementation of the Turner model (35), and is fully compatible with tertiary contacts such as pseudoknots, which pose difficulties for other tools (14, 57) but are however essential to model complex conformational changes such as those occurring in riboswitches.

State of the art design methods for RNA are based on computational frameworks to fold sequences on a given secondary structure from the knowledge of thermodynamics parameters for the pairing energies (74), and eventually including tertiary elements such as pseudoknots (81). Such methods have been used to obtain sequences with bistable secondary structures (22) and have been extended to take into account both positive and negative design elements (52, 81), and also to participative rational design (33).

The machine learning method implemented here includes two key ingredients differing with respect to the rational design: i) it exploits the sequences sampled through evolution and collected in sequence data bases, of SAM-I riboswitches sequence, building upon the frameworks introduced in homology and covariation detection (12, 42, 47, 52, 79); ii) it uses only the statistics of the natural sequence to build the model encoding at once in the parameter of the RBM model the multiple constraints which allow to the natural sequences in this family to properly fold in the secondary and tertiary structure and function by the allosteric response to SAM binding. Learning of the parameter of the RBM model is done in an unsupervised way through contrastive divergence (see Method Section) (27) which includes both a positive and negative design term.

We have first verified that the RBM model learned from sequence data encode structural constraints by predicting nucleotide-nucleotide contacts in the secondary structure and in the pseudoknot, performing at the same level of pairwise Potts models previously introduced to this aim ((12, 79). Deep network learning approaches recently introduced, goes further in the direction of structural predictions from sequence data aiming at a complete structure predictions and following the impressive success of analogous approaches in the protein field (29, 49, 51, 71).

We have then verified through SHAPE-probing that the design of synthetic bistable aptamers was successful for the RBM model but not for the CM model taking only into account secondary-structure constraints: 50% of artificial sequences with an average 20% distance with any natural one were switching conformation in response to SAM, while Infernal generated sequences were not.

To generate sequences with Infernal we have so far used the default parameters used for homology detection (47), called Rfam CM in the text. We have further extended the comparison with standard covariation and rational design models in three directions: i) better parametrizing the Infernal model for sequence generation to better reproduce the natural conservation profile in the generated sequences, ii) adding the pseudoknot in the covariance model, iii) generating sequences by rational design using the Infrared pipeline (81). For all the generated sequences in the three classes above the RBM energies are as large for the ones generated with the artificial Infernal sequences already tested, predicting that these sequences do not switch upon SAM binding (See Supplementary Figure S9). We plan to experimentally investigate this hypothesis in the future. Moreover we plan to extend our modeling of 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 can be applied to design other RNA from their respective MSA, including longer and more complex RNA such as ribosomal and messanger RNA.

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.

Conflict of interest statement. None declared.

REFERENCES

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. B. Bravi, A. Di Gioacchino, J. Fernandez-de Cossio-Diaz, A. M. Walczak, T. Mora, S. Cocco, and R. Monasson. Learning the differences: A

transfer-learning approach to predict antigen immunogenicity and T-cell receptor specificity. *bioRxiv*, pages 2022–12, 2022.

4. 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.
5. 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.
6. J. Chappell, M. K. Takahashi, and J. B. Lucks. Creating small transcription activating rnas. *Nature chemical biology*, 11(3):214–220, 2015.
7. 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.
8. 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.
9. S. Danisch and J. Krumbiegel. Makie.jl: Flexible high-performance data visualization for julia. *Journal of Open Source Software*, 6(65):3349, 2021.
10. K. Darty, A. Denise, and Y. Ponty. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25(15):1974, 2009.
11. 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.
12. 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.
13. 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.
14. 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.
15. G. Desjardins, H. Luo, A. Courville, and Y. Bengio. Deep tempering. *arXiv preprint arXiv:1410.0123*, 2014.
16. 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.
17. R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.
18. S. R. Eddy. Computational analysis of conserved rna secondary structure in transcriptomes and genomes. *Annual review of biophysics*, 43:433–456, 2014.
19. V. Epshtein, A. S. Mironov, and E. Nudler. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proceedings of the National Academy of Sciences*, 100(9):5052–5056, 2003.
20. M. P. Ferla and W. M. Patrick. Bacterial methionine biosynthesis. *Microbiology*, 160(8):1571–1584, 2014.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. G. E. Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.
28. 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.
29. 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.
30. 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.
31. 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.
32. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
33. 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.
34. W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.
35. 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.
36. 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.
37. D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
38. 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.
39. 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.
40. J. Melchior, A. Fischer, and L. Wiskott. How to center deep boltzmann machines. *The Journal of Machine Learning Research*, 17(1):3387–3447, 2016.
41. 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.
42. 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.
43. R. K. Montange and R. T. Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, 441(7097):1172–1175, 2006.
44. 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.
45. 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.
46. E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.
47. P. Nawrocki Eric and R. I. Eddy Sean. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.
48. R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

49. 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.
50. 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.
51. R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.
52. E. Rivas. Rna structure prediction using positive and negative evolutionary information. *PLoS computational biology*, 16(10):e1008387, 2020.
53. 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.
54. 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.
55. C. Roussel, J. Fernandez-de Cossio-Diaz, S. Cocco, and R. Monasson. Deep tempering with stacked restricted boltzmann machines. *HAL:03919483*, 2022.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. K. Shimagaki and M. Weigt. Selection of sequence motifs and generative hopfield-potts models for protein families. *Physical Review E*, 100(3):032128, 2019.
62. 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.
63. B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
64. C. D. Stoddard and R. T. Batey. Mix-and-match riboswitches, 2006.
65. 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.
66. 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.
67. Z. Sükösd, M. S. Swenson, J. Kjems, and C. E. Heitsch. Evaluating the accuracy of shape-directed rna secondary structure predictions. *Nucleic acids research*, 41(5):2807–2816, 2013.
68. D.-J. Tang, X. Du, Q. Shi, J.-L. Zhang, Y.-P. He, Y.-M. Chen, Z. Ming, D. Wang, W.-Y. Zhong, Y.-W. Liang, et al. A sam-i riboswitch with the ability to sense and respond to uncharged initiator trna. *Nature Communications*, 11(1):2794, 2020.
69. A. Tareen and J. B. Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020.
70. 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.
71. 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.
72. 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.
73. J. Tubiana, S. Cocco, and R. Monasson. Learning protein constitutive motifs from sequence data. *Elife*, 8:e39397, 2019.
74. 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.
75. 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.
76. J. X. Wang and R. R. Breaker. Riboswitches that sense s-adenosylmethionine and s-adenosylhomocysteine. *Biochemistry and Cell Biology*, 86(2):157–168, 2008.
77. 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.
78. 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.
79. 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.
80. W. C. Winkler, A. Nahvi, N. Sudarsan, J. E. Barrick, and R. R. Breaker. An mrna structure that controls gene expression by binding s-adenosylmethionine. *Nature Structural & Molecular Biology*, 10(9):701–707, 2003.
81. H.-T. Yao, Y. Ponty, and S. Will. Developing complex RNA design applications in the Infrared framework. In *RNA Folding - Methods and Protocols*. 2022.
82. M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of mathematical biology*, 46:591–621, 1984.