Nucleic Acids Research, 2024, 1–13
https://doi.org/10.1093/nar/gkae289
Computational Biology

Towards parsimonious generative modeling of RNA families

Francesco Calvanese[1,2], Camille N. Lambert[2], Philippe Nghe[2], Francesco Zamponi[3,4,*] and Martin Weigt[1,*]

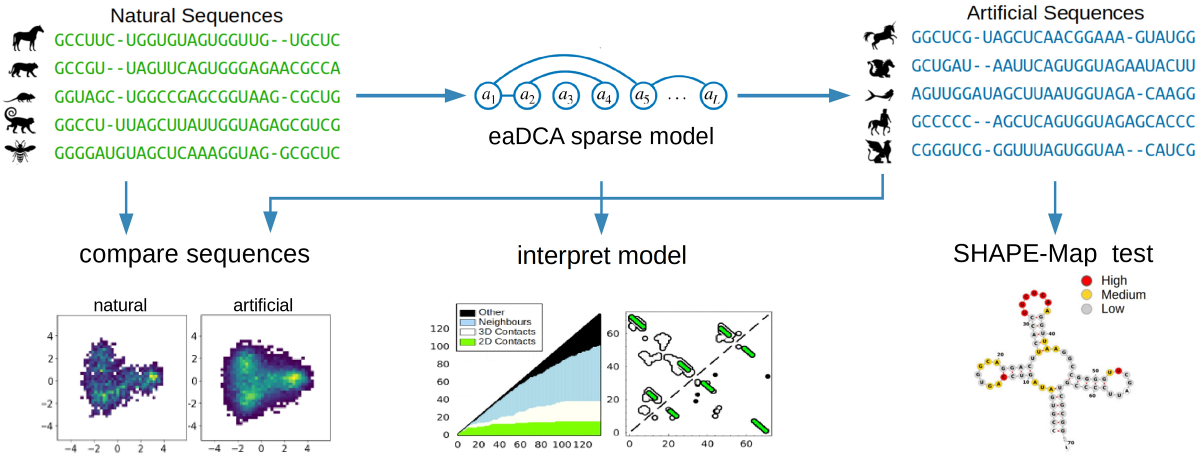
[1]Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratoire de Biologie Computationnelle et Quantitative – LCQB, Paris, France
[2]Laboratoire de Biophysique et Evolution, UMR CNRS-ESPCI 8231 Chimie Biologie Innovation, PSL University, Paris, France
[3]Dipartimento di Fisica, Sapienza Università di Roma, Rome, Italy
[4]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France

[*]To whom correspondence should be addressed. Email: martin.weigt@sorbonne-universite.fr
Correspondence may also be addressed to Francesco Zamponi. Email: francesco.zamponi@uniroma1.it

Abstract

Generative probabilistic models emerge as a new paradigm in data-driven, evolution-informed design of biomolecular sequences. This paper introduces a novel approach, called Edge Activation Direct Coupling Analysis (eaDCA), tailored to the characteristics of RNA sequences, with a strong emphasis on simplicity, efficiency, and interpretability. eaDCA explicitly constructs sparse coevolutionary models for RNA families, achieving performance levels comparable to more complex methods while utilizing a significantly lower number of parameters. Our approach demonstrates efficiency in generating artificial RNA sequences that closely resemble their natural counterparts in both statistical analyses and SHAPE-MaP experiments, and in predicting the effect of mutations. Notably, eaDCA provides a unique feature: estimating the number of potential functional sequences within a given RNA family. For example, in the case of cyclic di-AMP riboswitches (RF00379), our analysis suggests the existence of approximately 10^{39} functional nucleotide sequences. While huge compared to the known <4000 natural sequences, this number represents only a tiny fraction of the vast pool of nearly 10^{82} possible nucleotide sequences of the same length (136 nucleotides). These results underscore the promise of sparse and interpretable generative models, such as eaDCA, in enhancing our understanding of the expansive RNA sequence space.

Graphical abstract


Natural Sequences
GCCUUC-UGGUGUAGUGGUUG--UGCUC
GCCGU--UAGUUCAGUGGGAGAACGCCA
GGUAGC-UGGCCGAGCGGUAAG-CGCUG
GGCCU-UUAGCUUAUUGGUAGAGCGUCG
GGGGAUGUAGCUCAAAGGUAG-GCGCUC
a_1 a_2 a_3 a_4 a_5 … a_L
eaDCA sparse model
Artificial Sequences
GGCUCG-UAGCUCAACGGAAA-GUAUGG
GCUGAU--AAUUCAGUGGUAGAAUACUU
AGUUGGAUAGCUUAAUGGUAGA-CAAGG
GCCCCC--AGCUCAGUGGUAGAGCACCC
CGGGUCG-GGUUUAGUGGUAA--CAUCG
compare sequences
natural
artificial
interpret model
Other
Neighbours
3D Contacts
2D Contacts
SHAPE-Map test
High
Medium
Low

Introduction

RNA molecules play a critical role in many biological processes, including gene expression and regulation. They carry a multitude of functions, such as encoding and transferring genetic information, regulating gene expression and catalyzing chemical reactions (1–3). Functional RNA molecules are expected to be extremely rare in the exponentially vast nucleotide-sequence space, and current databases contain only a tiny fraction of the overall possible, functionally viable sequence diversity. However, it is worth noting that almost identical biological functions can be carried out by different RNA exhibiting substantial sequence variability. Databases like Rfam (4) gather these in diverse yet functionally consistent families of homologous RNA sequences. In computational se-

Received: October 14, 2023. **Revised:** March 5, 2024. **Editorial Decision:** March 31, 2024. **Accepted:** April 5, 2024

quence biology, a significant challenge lies in harnessing the relatively limited pool of existing RNA sequences within a family, often comprising just a few hundred or thousand examples. The objective is to decipher the sequence patterns that underpin the three-dimensional structure and biological functions of these RNA families. This endeavor extends beyond the known sequences, aiming to explore the vast potential space of sequences capable of adopting similar structures and functions. Such analyses provide valuable insights into the complex organization of sequence space and, ultimately, unravel the intricate sequence-to-function relationship. This quest has gained paramount significance, especially in the era of high-throughput sequencing, solidifying its status as one of biology's central and most challenging questions.

Generative probabilistic models offer a powerful approach to tackling these challenges by extrapolating beyond the limited pool of known RNA molecules and generating previously unseen functional sequences. When applied to RNA families, these models build a probability distribution, denoted as $P(a_1, \ldots, a_L)$ (5–11). This distribution encapsulates the variability found in the known sequences within the family while encompassing all possible sequences of length L (for a more precise definition, see *Material and Methods*). To provide an intuitive analogy, think of this probability distribution as defining a 'landscape' across sequence space assigning high probabilities to functional sequences, akin to the peaks in this landscape. Conversely, non-functional sequences receive low probabilities.

This probability distribution also enables the prediction of mutational effects (12,13) since mutations can alter the sequence probabilities relative to the wildtype. Additionally, these models allow for generating novel synthetic sequences (9,14) through a sampling process (as illustrated in Figure 1). A well-constructed model P should possess the ability to generate nucleotide sequences that are diverse but statistically indistinguishable from the known sequences in the family.

An RNA generative model $P(a_1, \ldots, a_L)$ has to assign probabilities to a huge number of potential sequences $(a_1, \ldots, a_L)$ while learning from a relatively small pool of observed sequences. As an example, consider an RNA molecule with an aligned length of $L = 150$ residues, i.e. the sequence may contain both nucleotides and gaps. The model must provide up to $5^L \simeq 10^{105}$ probabilities, even though typical RNA families consist of only $10^2 - 10^4$ known sequences, cf. (4). The lack of abundant RNA data makes it hard for complex models like deep-learning architectures to work well, as seen in other tasks like RNA structure prediction (15). This suggests that simpler, less complex models may be better suited to tackle RNA. The currently most successful class of probabilistic sequence models are covariance models (CM) (10,11); they model conservation both in nucleotide sequence and of secondary structure (resulting in covariation of paired nucleotides) across families of homologous RNA. Being sensitive and efficient computational tools for RNA homology search and alignment, they form the methodological basis for the construction of the Rfam database (4). The Boltzmann Machine Direct Coupling Analysis (bmDCA) (16,17) models covariation also for nucleotides not paired in the secondary structure.

The core idea behind CM and bmDCA lies in the notion that RNA residues of significant functional importance experience evolutionary pressures that deter deleterious mutations. Consequently, these residues tend to remain conserved across the Multiple Sequence Alignment (MSA) collecting ex-

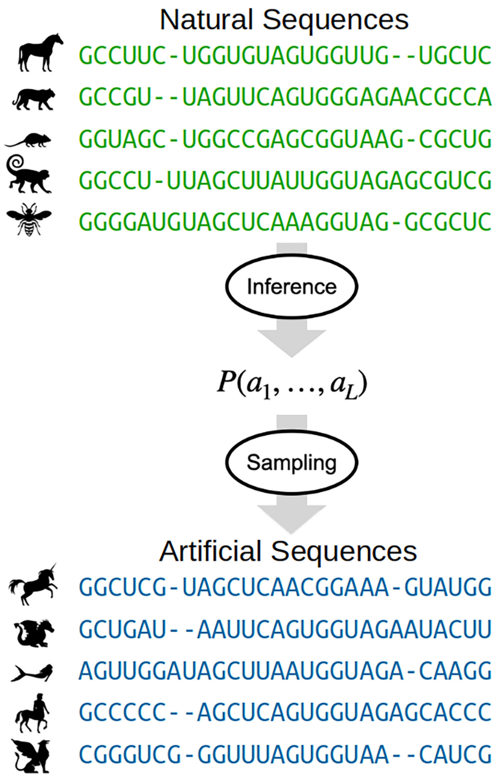

Natural Sequences
GCCUUC-UGGUGUAGUGGUUG--UGCUC
GCCGU--UAGUUCAGUGGGAGAACGCCA
GGUAGC-UGGCCGAGCGGUAAG-CGCUG
GGCCU-UUAGCUUAUUGGUAGAGCGUCG
GGGGAUGUAGCUCAAAGGUAG-GCGCUC
Inference
$P(a_1, \ldots, a_L)$
Sampling
Artificial Sequences
GGCUCG-UAGCUCAACGGAAA-GUAUGG
GCUGAU--AAUUCAGUGGUAGAAUACUU
AGUUGGAUAGCUUAAUGGUAGA-CAAGG
GCCCCC--AGCUCAGUGGUAGAGCACCC
CGGGUCG-GGUUUAGUGGUAA--CAUCG

Figure 1. Probabilistic generative models extract a probability distribution for the RNA family from natural data, which can then be used to generate artificial sequences. The generated sequences are statistically similar to the natural ones, yet they differ from any existing variant, thereby introducing an element of novelty.

tant homologous sequences. Conversely, pairs of nucleotides that exhibit co-evolutionary patterns over time display correlated mutations. To capture both types of constraints, CM and bmDCA adjust its probability distribution to mirror the one-site and two-site frequencies observed in the MSA, which serve as proxies for conservation and co-evolution. As distinctive features between the two approaches, CM restrict modeled coevolution to secondary-structure pairings, but it scores also insertions and deletions, making it applicable to unaligned sequences. bmDCA models coevolution between all nucleotide pairs, cf. reviews in (18,19), but it requires aligned sequences as an input. The resulting RNA generative models were observed to be accurate (8), but their fully connected graphical structure limits computational efficiency and biological interpretability.

In this context, one-site frequencies, denoted as $f_i(a)$, describe how often a nucleotide $a \in \{A, U, C, G, -\}$ (with '−' representing alignment gaps) appears at a specific site $i \in 1, \ldots, L$ within the MSA. Meanwhile, two-site frequencies, denoted as $f_{ij}(a, b)$, provide information about the joint occurrence of nucleotide pairs (a, b) at positions (i, j) within the same sequence. The probability distribution used in bmDCA takes the form of a fully connected Potts model/Markov Random Field, which captures the interplay of these frequencies,

$$P(a_1, \ldots, a_L) = \frac{1}{Z} \exp\left\{ \sum_{i=1}^{L} h_i(a_i) + \sum_{i<j} J_{ij}(a_i, a_j) \right\}, \quad (1)$$

with Z being the partition function that guarantees normalization. The $h_i(a)$ $(a \in \{A, U, C, G\ -\})$ are the local 'fields' used to fit the one-site statistics. The $J_{ij}(a, b)$ matrices (with $(a, b) \in \{A, U, C, G, -\}^2$) are 5×5 interaction 'couplings' used to fit the two-site statistics.

Although DCA has proven itself as a valuable instrument in investigating proteins, exhibiting achievements in tasks like generating functional sequences (14), forecasting the effect of mutations (12,13), deciphering protein evolution (20,21), and identifying structural interactions (22,23), its application to RNA remains relatively unexplored (5–8). Furthermore, the limited availability of RNA data, compared to the wealth of data for proteins, makes the use of intricate models like large language models (24) impractical. Consequently, employing simpler models for RNA is not only suitable but also presents the benefits of enhanced interpretability, reduced computational burden, and local trainability.

Nonetheless, conventional bmDCA generates a fully connected coupling network (as seen in Eq. (1)): it models coevolution between all conceivable pairs of residues, even when there is no actual coevolution occurring. As a consequence, this approach can yield a substantial number of noisy couplings $J_{ij}(a, b)$ in the network that lack any statistical support. To mitigate this issue, network sparsification can be applied to trim down the network by eliminating numerous spurious couplings. This process aids in identifying the most informative and functionally significant couplings, rendering the network more accessible for interpretation and analysis. Previous endeavors in this direction have primarily concentrated on sparsifying coupling networks within proteins (25).

In our work, we introduce a novel approach called Edge Activation Direct Coupling Analysis (eaDCA). Unlike previous algorithms, eaDCA takes a unique starting point: an empty coupling network. It then systematically constructs a non-trivial network from scratch, rather than starting with a fully connected network like bmDCA and subsequently simplifying it, or using external information like secondary structure in CM. Our step-by-step process generates a series of models, gradually increasing in complexity until they achieve a statistical performance comparable to that of bmDCA.

Our method offers some notable advantages: first, it results in generative Potts models with considerably less parameters compared to the standard bmDCA, activating couplings only between nucleotide pairs showing direct coevolution. Second, the algorithm is substantially more efficient than bmDCA, greatly reducing the time required for model learning. Third, at each stage of our approach, we employ analytical likelihood maximization. This allows us to easily track normalized sequence probabilities and estimate entropies throughout the network-building process. This valuable information enhances our ability to interpret and analyze the vast space of RNA sequences.

The organization of the manuscript is as follows. In 'Materials and Methods', we present the foundational principles and functionality of the model, describe the data used in the model training and analysis, and provide specific information about the SHAPE-MaP experiments conducted to examine artificial molecules. In 'Results and Discussion', we evaluate the statistical properties of the artificial sequences generated by eaDCA, interpret the parameters of the sparse architectures, and examine the model's predictions regarding mutational effects on tRNA. Additionally, using eaDCA to access normalized sequence probabilities and model entropies, we conduct an analysis on how different constraints, such as compatibility with secondary structures or family conservation and coevolution statistics, affect the size of the viable RNA sequence space. Lastly, we assess the SHAPE-MaP experimental results, characterizing the structure of artificially generated tRNA molecules.

Materials and methods

In this section, we discuss the data and the methodological basis of our work: all data used for training and evaluating our models, the new algorithm proposed here, and the experimental protocol to test artificial sequences generated by our approach.

Data

RNA families

All generative models discussed here are trained for individual RNA families, i.e. homologous but diverged sequences of largely conserved structure and function (4). Each family is represented by a Multiple Sequence Alignment (MSA) $\mathcal{D} = (a_i^r | i = 1, \ldots, L; r = 1, \ldots, M)$, with L indicating the aligned sequence length, and M the number of distinct sequences. The entries a_i^r are either one of the four nucleotides $\{A, U, C, G\}$, or the alignment gap '–' reflecting insertions and deletions in the original unaligned sequences.

Following standards in the literature, phylogenetic effects are partially compensated by reweighting each sequence by a factor ω_r (19), which equals the inverse number of all sequences having more than 80% sequence identity to sequence r, and which is used when estimating the empirical single-site nucleotide frequencies $f_i(a)$ and pair frequencies $f_{ij}(a, b)$ from the data $\mathcal{D}$, cf. the supplementary information (SI) for details. The sum of weights $M_{\text{eff}} = \sum_r \omega_r$ defines the effective sequence number as a more accurate reflection of the diversity of the dataset.

eaDCA is tested on 25 RNA families of known tertiary structure with L ranging from 50 to 350 and M from 30 to 50000. These families are extracted from the CoCoNet benchmark dataset (26) by limiting ourselves to datasets with high M_{eff} and sequence length $L < 350$. The MSA were updated using a more recent Rfam release (May 2022), and matched to exemplary PDB structures. A comprehensive list of family names, characteristics, and used PDBs is given in Supplementary Section S1, Supplementary Table S1.

The main text concentrates on two families: the tRNA family (RF00005) was selected due to the existence of mutational datasets, and our own experiments were performed on this family. Due to its unusually large size, the MSA was randomly downsampled to $M = 30\,000$ sequences. The cyclic di-AMP riboswitches (RF00379) were chosen due to their interesting and non-trivial statistical properties. The robustness of all results is illustrated in the SI, where the other 23 families are exhaustively analyzed.

Mutational fitness dataset

To evaluate our ability to predict mutational effects in RNA molecules, we utilized the data published in (27). This dataset provides *in vivo* fitness measurements for 23,283 variants of the yeast $\text{tRNA}^{\text{CCU}}_{\text{Arg}}$ at temperatures of 23, 30 and 37°C, with up to 10 mutations compared to wildtype. These mutations may result in non-functional sequence variants, in difference to the natural sequences in the RNA families. We focus on the results at 37°C because, at higher temperature, the $\text{tRNA}^{\text{CCU}}_{\text{Arg}}$

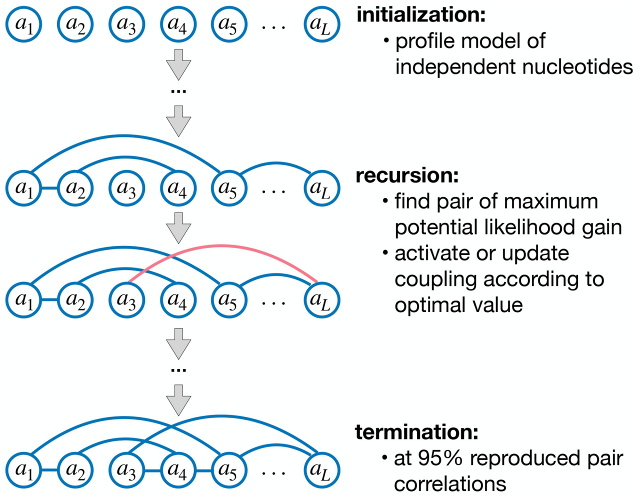

a_1 a_2 a_3 a_4 a_5 … a_L
initialization:
• profile model of independent nucleotides
...
recursion:
• find pair of maximum potential likelihood gain
• activate or update coupling according to optimal value
...
termination:
• at 95% reproduced pair correlations

Figure 2. Schematic representation of the recursive eaDCA algorithm.

becomes increasingly important for the survival of the organism. Fitness values in (27) are organized such that 0.5 represents a mutant yeast strain incapable of reproduction, while 1.0 is the wildtype fitness. The details of the datasets, the fitness definition of (27), and our results for 23 and 30°C are provided in the SI

SHAPE reference dataset

In order to empirically validate our generative models, we conducted Selective 2’-Hydroxyl Acylation analyzed by Primer Extension with Mutational Profiling (SHAPE-MaP) experiments on artificially generated tRNA molecules, cf. below. To ensure the robustness of our analysis and to facilitate a meaningful comparison, we utilized an external published dataset comprising SHAPE reactivity profiles for 20 RNA sequences with known secondary structure. This dataset, which we will refer to as the ‘SHAPE Reference Dataset’, was obtained from (28) .

Edge activation direct coupling analysis (eaDCA)

Algorithm principle

The proposed algorithm belongs to the family of DCA algorithms, i.e. it learns a Potts model in the form of Eq. (1) from an MSA $\mathcal{D}$. However, instead of introducing couplings $J_{ij}(a, b)$ for all pairs of nucleotide positions $0 \le i < j \le L$, we aim at a parsimonious model and activate couplings only between those pairs, which are really coevolving and thus essential for the accurate statistical description of the sequence family. All other pairs, which do not have clear signatures of direct coevolution, shall not be included into the set of coevolutionary couplings, to avoid noise overfitting (25) .

Since the empirical pair frequencies $f_{ij}(a, b)$ are shaped both by direct coupling and indirect correlation, the set of coupled pairs, $\mathcal{E} = \{(ij) \mid J_{ij}$ is non-zero$\}$, cannot be fixed in a single step, but has to be constructed recursively, as is shown schematically in Figure 2 and detailed below: starting from a profile model of independent nucleotides, $\mathcal{E}_0 = \emptyset$, we construct a series of edge sets $\mathcal{E}_t$, by activating or updating edges one by one. In this setting, ‘activating’ an edge signifies to introduce a non-zero coupling for a previously uncoupled pair (ij), while ‘updating’ indicates a change of the coupling value on an already activated edge. As a consequence, at any algorithmic step t, the model can be written as

$$P_t(a_1, \ldots, a_L) = \frac{1}{Z_t} \exp\{-E_t(a_1, \ldots, a_L)\}$$

$$E_t(a_1, \ldots, a_L) = -\sum_{i=1}^{L} h_i(a_i) - \sum_{(ij)\in\mathcal{E}_t} J_{ij}(a_i, a_j), \tag{2}$$

with E_t being called ‘statistical energy’. The log-likelihood of the model given the reweighted data $\mathcal{D}$ reads

$$\mathcal{L}_t = \sum_{r=1}^{M} \omega_r \log P_t(a_1^r, \ldots, a_L^r). \tag{3}$$

Initialization

As already mentioned, the model is initialized without couplings, $\mathcal{E}_0 = \emptyset$, and reads

$$P_0(a_1, \ldots, a_L) = \frac{1}{Z_0} \exp\left\{\sum_{i=1}^{L} h_i(a_i)\right\}. \tag{4}$$

The log-likelihood $\mathcal{L}_0$ is easily maximized by setting

$$h_i(a) = \log f_i(a) \tag{5}$$

for all $i = 1, \ldots, L$ and $a \in \{A, U, C, G, -\}$, i.e. the model reproduces the empirical single-residue statistics. The resulting partition function is $Z_0 = 1$. This simple model is known under the name of profile model (or independent-site model) and widely used in bioinformatic sequence analysis.

Recursion

The algorithmic step from t to $t + 1$ is characterized by a modification of a single 5×5 coupling matrix J_{kl} on a single position pair (kl),

$$E_{t+1}(a_1, \ldots, a_L) = E_t(a_1, \ldots, a_L) - \Delta J^*_{kl}(a_k, a_l),$$
$$\mathcal{E}_{t+1} = \mathcal{E}_t \cup \{(kl)\}. \tag{6}$$

If (kl) was not yet active in $\mathcal{E}_t$, this corresponds to an edge activation, otherwise to an edge update.

The edge (kl) and the coupling change $\Delta J^*_{kl}(a_k, a_l)$ are chosen to maximize the log-likelihood $\mathcal{L}_{t+1}$. As is proven in Supplementary Section S2.1, this is realized by (1) choosing the ‘least accurate’ position pair

$$(kl) = \underset{1\le m<n\le L}{\operatorname{argmax}} D_{KL}(f_{mn} \| P^t_{mn}), \tag{7}$$

which maximizes, over all possible position pairs (mn), the Kullback–Leibler divergence,

$$D_{KL}\left(f_{mn} || P^t_{mn}\right) = \sum_{a,b} f_{mn}(a, b) \log \frac{f_{mn}(a, b)}{P^t_{mn}(a, b)}, \tag{8}$$

between the empirical target distribution f_{mn} and the current model’s marginal two-site distribution P^t_{mn} defined as

$$P^t_{mn}(a, b) = \sum_{a_1,\ldots,a_L} P^t(a_1, \ldots, a_L)\delta_{a,a_m}\delta_{b,a_n}, \tag{9}$$

and by (2) activating/updating the coupling on the chosen position pair via

$$\Delta J^*_{kl}(a, b) = \log \frac{f_{kl}(a, b)}{P^t_{kl}(a, b)}. \tag{10}$$

To avoid excessively high values for rare nucleotide combinations, this coupling term is regularized using pseudocounts for both the empirical frequencies and the model probabilities, cf. Supplementary Section S2.2 for details.

Note that the selection goes over all position pairs (mn), independently on their activation status in $\mathcal{E}_t$. Note also that the exact determination of the marginal distributions $P^t_{mn}(a, b)$ in Eq. (9) is infeasible since it would require to sum over all 5^L possible sequences of aligned length L. We therefore use Markov chain Monte Carlo (MCMC) sampling; the exact procedure based on persistent contrastive divergence is detailed in Supplementary Section S2.3.

Termination

As the process continues, the resulting models become increasingly accurate and complex. It can be observed from Eq. (10) that a fixed point is attained when all the two-point probabilities are equal to their respective empirical frequencies (and in consequence also all single-site frequencies). This corresponds exactly to the fixed-point condition imposed in bmDCA. Because this condition is impossible to achieve in practice due to MCMC sampling noise, we set an *ad hoc* stopping criterion by looking at how well the empirical two-site covariances

$$c_{ij}(a, b) = f_{ij}(a, b) - f_i(a) f_j(b) \tag{11}$$

are reproduced by the connected correlations in the model,

$$c^t_{ij}(a, b) = P^t_{ij}(a, b) - P^t_i(a) P^t_j(b). \tag{12}$$

The algorithm terminates at step t_f when the Pearson correlation ρ between these two quantities, evaluated over all positions (ij) (including those not in $\mathcal{E}_{t_f}$) and all nucleotides a, b (including gaps), reaches 0.95. This value is commonly reached in bmDCA as well. The reason for computing the score based on the $c_{ij}(a, b)$ instead of the $f_{ij}(a, b)$ is that the former isolates coevolution statistics from the conservation ones.

The entire procedure is summarized as a pseudocode in Algorithm 1, and represented graphically in Figure 2.

Algorithm 1 (eaDCA)

```
Initialization:
    • Profile model: $P_0(a_1,\ldots,a_L) = \prod_{i=1}^{L} f_i(a_i)$
    • Iteration counter: $t \leftarrow 0$
Recursion:
While $\rho(c^t_{ij}, c_{ij}) < 0.95$ do
    • Estimate two-point probabilities $P^t_{ij}(a,b)$ for all $i,j$ and all $a,b$ via MCMC sampling
    • Identify the worst represented edge $(kl)$ according to Eq. (7)
    • Update the interaction on the identified edge using Eq. (10) to get the new model $P_{t+1}(a_1,\ldots,a_L)$
    • Add the identified edge $(kl)$ to $\mathcal{E}_t$ to get $\mathcal{E}_{t+1}$
    • Increment iteration counter: $t \leftarrow t+1$
Termination at $t_f = t$ :
    • Output $P_{t_f}(a_1,\ldots,a_L)$ with 95% reproduced pair correlations
```

A typical run of the training process is described in Supplementary Section S2.4 and Supplementary Figure S1. The reduction of the running time over bmDCA is exemplified in Section S2.5, Table S2.

Note that the eaDCA algorithm has similar objectives as the adaptive cluster expansion (ACE) proposed before to learn sparse Potts models from data (29,30). However, ACE requires an exact calculation of cluster partition functions, which limits its practical applicability to relatively small clusters of sequence positions. On the contrary, eaDCA is based on sampling which, even if introducing some level of stochasticity, allows for treating arbitrary interaction graphs ranging from very sparse to fully connected ones, in function of what is needed to capture the data statistics.

Normalized sequence probabilities and model entropy

Probabilistic generative models, including bmDCA, typically do not provide normalized sequence probabilities but only relative sequence weights. This limitation arises because obtaining normalized probabilities would necessitate summing over the entire 5^L sequence space to get the partition function Z given by Eq. (1),

$$Z = \sum_{a_1,\ldots,a_L} \exp\left\{ \sum_{i=1}^{L} h_i(a_i) + \sum_{(ij)\in\mathcal{E}} J_{ij}(a_i, a_j) \right\}, \tag{13}$$

which is infeasible for any biologically relevant value of L. Relative weights are sufficient for MCMC sampling of artificial sequences, but they are meaningful just within the context of a specific model, and cannot be compared across distinct models.

The advantage of eaDCA is that the recursion preserves model's partition function Z, as is shown in Supplementary Section S2.6. Since P_0 is trivially normalized, we have

$$Z_0 = 1 \text{ and } Z_{t+1} = Z_t, \tag{14}$$

i.e. the models remain trivially normalized under recursion:

$$P_t(a_1, \ldots, a_L) = \exp\{-E_t(a_1, \ldots, a_L)\}. \tag{15}$$

A nice consequence of this property is that we have easy access to the model's entropy S_t

$$\begin{aligned} S_t &= -\langle \log P_t(a_1, \ldots, a_L) \rangle_{P_t} \\ &= \langle E_t(a_1, \ldots, a_L) \rangle_{P_t} \end{aligned} \tag{16}$$

via the average statistical energy, which can be accurately estimated from an MCMC sample. From the entropy S_t we can deduce the size of the viable sequence space,

$$\Omega_t = \exp\{S_t\}, \tag{17}$$

which can be thought of as the effective number of different sequences that we can sample from $P_t(a_1, \ldots, a_L)$.

In practice, because we depend on stochastic MCMC techniques for estimating the two-site probabilities $P_{ij}(a, b)$ in eaDCA iterations, the Z is only approximately conserved. However, it is straightforward to accurately monitor and account for these errors, see the Supplementary Section S2.6 and Supplementary Table S3 for details.

SHAPE-MaP probing of artificial tRNA molecules

To conduct an empirical evaluation of our eaDCA-derived model, we performed a SHAPE-MaP analysis on a set of 76 artificially generated tRNA molecules (RF00005 family). Here we summarize the experimental protocols, full details are provided in the SI.

RNA production

We designed a total of 76 tRNAs. Each RNA was synthesized with the T7 promoter positioned at its 5' end and the last

16 nucleotides were kept constant for all constructs matching those of the yeast tRNA(asp). The DNA templates (gBlock or oligoPools from Integrated DNA Technologies) were amplified by PCR using the Phusion Hot Start Flex polymerase (New England Biolab). After purification, the DNAs were transcribed via *in vitro* transcription using the HiScribe T7 High Yield RNA Synthesis Kit (NEB). The resulting RNAs were purified by denaturing gel electrophoresis.

RNA modification

The SHAPE reactivity is not only a reflection of RNA structure but also depends on experimental conditions, necessitating careful consideration in comparative analysis of SHAPE-MaP reactivity profiles (31) . Consequently, we chose to probe our artificial tRNA with the same folding buffer (50 mM HEPES pH 8.0, 200 mM potassium acetate pH 8.0, and 3 mM $MgCl_2$) than the yeast tRNA(asp) Reference SHAPE Dataset (28) . For RNA modification, three conditions were performed: positive (with the probe), negative (only the probe solvent) and denaturing (denatured RNA with the probe). For the positive and negative conditions, the RNAs were allowed to refold and the modifying agent (1M7 in DMSO for positive) or the solvent (neat DMSO for negative) were quickly mixed to the RNAs and incubated 5 min at 37°C. For the denaturing condition, the RNAs were first denatured by addition of formamide followed by a heat treatment and the RNAs were modified similarly (1M7 probe). After incubation, all modified RNAs were purified via ethanol precipitation and quantified by the Qubit RNA High Sensitivity assay kit (ThermoFisher).

Library preparation

The modified RNAs were pooled in equimolar proportion based on their conditions (positive, negative, denaturing) and reverse-transcribed using the SuperScript II reverse transcriptase (ThermoFisher) with a buffer allowing the misincorporation of nucleotides at the chemically modified positions. We also used a Template Switching Oligo (TSO) in order to incorporate the Rd1 Illumina adapter during the reverse-transcription, and brought the Rd2 Illumina adapter by the reverse-transcription primer. After cDNA purification, PCR enrichment was conducted to amplify the DNA libraries and incorporate the P5/P7 Illumina adaptors. The samples were purified by AMPure XP beads (Beckman Coulter), quantified by quantitative PCR (KAPA Library Quantification Kit, Roche), and sequenced on a MiSeq-V3 flow cell (Illumina) at the NGS platform of Institut Curie (Paris, France). For in depth details about the procedure and the primers used refer to SI.

SHAPE reactivity mapping

We employed the ShapeMapper2 (32) software to process the sequencing data, obtaining SHAPE reactivity values for each artificial tRNA molecule, which partition sites into the reactivity classes ‘low’, ‘medium’ and ‘high’ (31,33). ShapeMapper2 was run with default settings, except for the depth-per-site quality threshold that was lowered from 5000 to 3000. This allowed us to gather reactivity data covering more than 50% or the residues for at least 30 of the molecules under investigation.

Results and discussion

eaDCA models reproduce the natural sequence statistics

The initial evaluation of the performance of any generative model involves assessing its ability to accurately replicate the statistical properties of natural data. To this aim, we conduct an analysis across 25 RNA families, and compare the statistical properties of the natural sequences from the family’s MSA with those of a large number of artificial sequences, which are independently sampled from the eaDCA model $P(a_1, ..., a_L)$. For comparison, we also sample from a simpler, secondary-structure based model (SSBM), where only nucleotide pairs involved in the secondary structure (S2D) are connected by couplings. The SSBM is also a Potts model, with all maximum-likelihood parameters derived exactly from the empirical one- and two-nucleotide statistics,

$$P_{SSBM}(a_1,\ldots,a_L) = \exp\left\{\sum_{i=1}^{L} h_i(a_i) + \sum_{(ij)\in S2D} K_{ij}(a_i,a_j)\right\}$$

$$K_{ij}(a_i,a_j) = \log\frac{f_{ij}(a_i,a_j)}{f_i(a_i)f_j(a_j)}, \quad h_i(a_i) = \log f_i(a_i). \tag{18}$$

We use SSBM as a performance benchmark because the information about RNA secondary structure is readily available for all Rfam families, and because base-pair complementarity in RNA secondary structure causes a strong pairwise coevolution. Note that these models bare similarities to, but are different from the CM used in Infernal (10,11), since they do not model insertions and deletions, and thus require already aligned input sequences.

Figure 3 displays statistical analyses for the RF00005 and RF00379 families. Additional results for 23 other RNA families can be found Supplementary Figure S2. Figures 3D and 3H display the comparison between the connected two-point correlations of the natural data with estimates from a sample of an eaDCA obtained model and the SSBM. The results indicate a strong correlation of eaDCA with the natural data for all residue pairs, including those not connected by activated edges, while the SSBM reproduces the pair correlations only on the secondary structure, and totally fails on all other pairs of positions.

A second test of the eaDCA model’s generative properties is demonstrated in Figures 3A–C, and E–G, which present the natural, eaDCA, and SSBM generated sequences projected onto the first two principal components (PCs) of the natural MSA (14,34). The sequences sampled from the eaDCA model effectively reproduce the visible clustered structure of the natural sequences, while SSBM are unable to do so, with projections on the PCs being concentrated around the origin.

The observations in Figure 3 indicate the inability of SSBM to serve as accurate generative models, while sequences sampled from eaDCA are coherent with the natural data on the tested observables. This suggests that the Potts model requires more than just the secondary-structure based interaction couplings to function properly and that an overly aggressive reduction in parameters compromises the model’s performance. Sequences directly emitted from CM (`cmemit` command of Infernal) are analyzed in Supplementary Figure S3: CM slightly outperform the simpler SSBM, but remain less accurate than eaDCA.

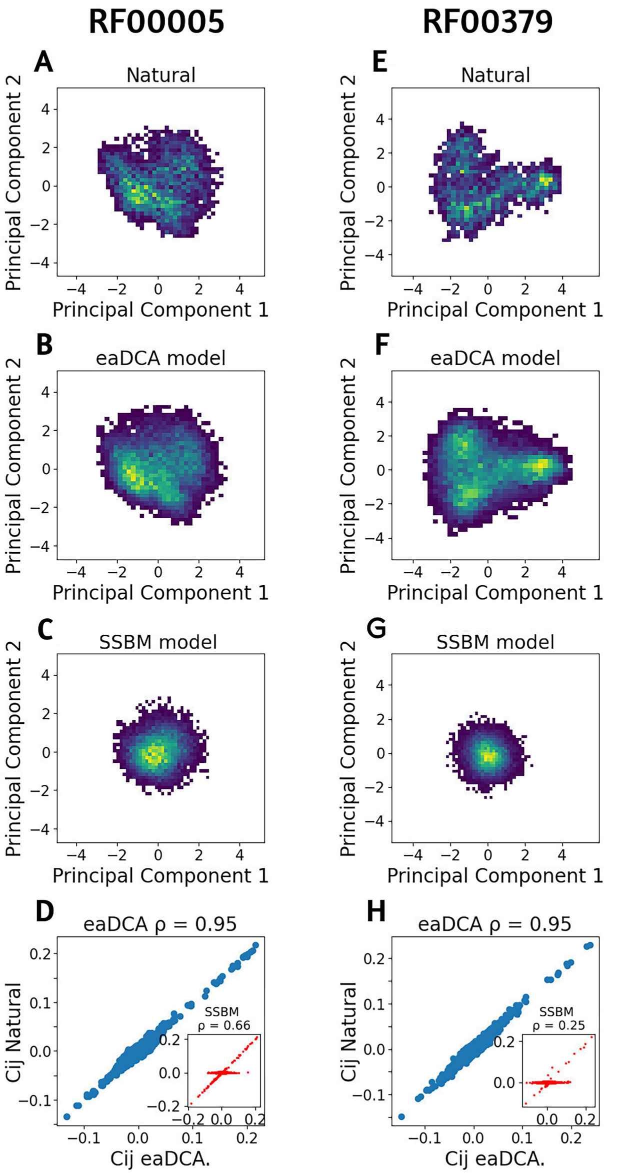


Figure 3. (**A**) RF0005: principal component analysis of natural sequences ($M = 28770$). (**B**) RF005: eaDCA generated sequences mapped to the first two principal components of the natural sequences ($M = 12000$). (**C**) RF005: SSBM generated sequences mapped to the first two principal components of the natural sequences ($N = 12\,000$). (**D**) RF00005: scatter plot of the connected two-site correlations of the natural sequences vs. eaDCA generated sequences (blue) or SSBM generated sequences (red - insert) ($N = 12\,000$). (**E**) RF00379: Principal component analysis of natural sequences ($N = 3808$). (**F**) RF00379: eaDCA generated sequences mapped to the first two principal components of the natural sequences ($N = 12000$). (**G**) RF00379: SSBM generated sequences mapped to the first two principal components of the natural sequences ($N = 12\,000$). (**H**) RF00379: scatter plot of the connected two-site correlations of the natural sequences vs. eaDCA generated sequences (blue) or SSBM generated sequences (red: insert) ($N = 12\,000$).

From Table 1, we conclude that eaDCA delivers generative models able to reproduce the natural RNA statistics with only a fraction of the number of parameters of a standard bmDCA implementation (parameter reduction of 84.85% for RF00005 and 87.83% for RF00379). A complete table for all the 25 families is included in Supplementary Table S1 and confirms this observation across families.

Parameter interpretation

A key benefit of employing a parsimonious generative model is the potential for obtaining a more insightful interpretation of its parameters. In the context of RNA, the eaDCA method is producing good generative models with a small percentage of the parameters of fully connected models (bmDCA), which in turn enables easier biological interpretation. Since the edge activation procedure starts from the profile model, all single-site frequencies $f_i(a)$ are accurately reproduced from the beginning. Due to its iterative nature, eaDCA produces additionally an ordered list of edges carrying non-zero couplings. These added edges can be used to explain the connected two-point statistics to high accuracy, and they are thus carrying the full information about residue coevolution in the MSA of the RNA family under consideration.

For this study, we classified the first L added edges into four categories: 'secondary structure base pairs' ($S2D$), 'neighbors' (if the pair is less than four positions apart along the primary sequence), 'tertiary structure contacts' (if the distance between the involved residues is less than 8Å, but the pair is neither S2D nor neighboring), and 'other' (not fitting into any of the prior categories). We present here the analyses for two RNA families (RF0005 and RF00379) but the results of all 25 families can be found in Supplementary Figure S4.

In Figure 4, the analysis revealed a relationship between contacts and added edges, with almost all $S2D$ pairs being systematically taken in the early iterations. This trend is consistent with their strong coevolutionary relationship, and shows that SSBM and CM models capture many of the strongest, but by far not all such relationships. Tertiary contacts are included later (and many never activated even at termination of the algorithm); we therefore conclude that they typically induce a much lower coevolutionary signal than secondary-structure contacts. The presence of activated edges between neighboring residues may in part be attributable to phylogenetic relationships, but also to the insertion or deletion of multiple nucleotides, i.e. to the presence of gap stretches in the MSA.

A relatively small fraction of activated edges do not offer an interpretation (class 'other'), it remains unclear if these edges reflect the limited statistics in the natural MSAs, or coevolution beyond structural contacts. eaDCA considers them important for reproducing the natural sequence statistics. In this context, it is important to note that the complete list of edges generated by eaDCA before meeting the termination condition significantly exceeds the sequence length L, consequently leading to a large quantity of 'other' entries.

These observations may suggest to use eaDCA for RNA structure prediction. Since, however, the focus of our work are generative probabilistic models rather than structure prediction, our restrict this Section to interpretative analysis. Actually, the natural MSA are constructed using the `cmbuild` and `cmalign` commands from Infernal (11), which actively incorporate a given secondary structure, and therefore may bias applications to structure prediction. However, to achieve a more comprehensive picture, we have included in Supplementary Figure S5 and Supplementary Table S4 a comparison of the contact retrieval capability of eaDCA against Evolutionary Couplings (5,6) and R-scape (35) (trained on the same alignments).

Prediction of mutational effects

Potts models (including profile, SSBM and DCA models) are energy-based statistical model, cf. Eq. (1). The maximum-

Table 1. eaDCA and SSBM results for RF00005 (tRNA) and RF00379 (cyclic di-AMP riboswitch) at termination $t = t_f$. PR% indicates the percentage of parameter reduction, S the eaDCA model entropy, and Ω the corresponding effectoive size of viable sequence space

Rfam Id	L	M	M_{eff}	SSBM c_{ij} corr	eaDCA c_{ij} corr	PR%*	S	Ω
RF00005	71	28770	2267	0.66	0.95	84.85%	51.34	1.98×10^{22}
RF00379	136	3808	1428	0.25	0.95	87.83%	89.56	1.05×10^{39}

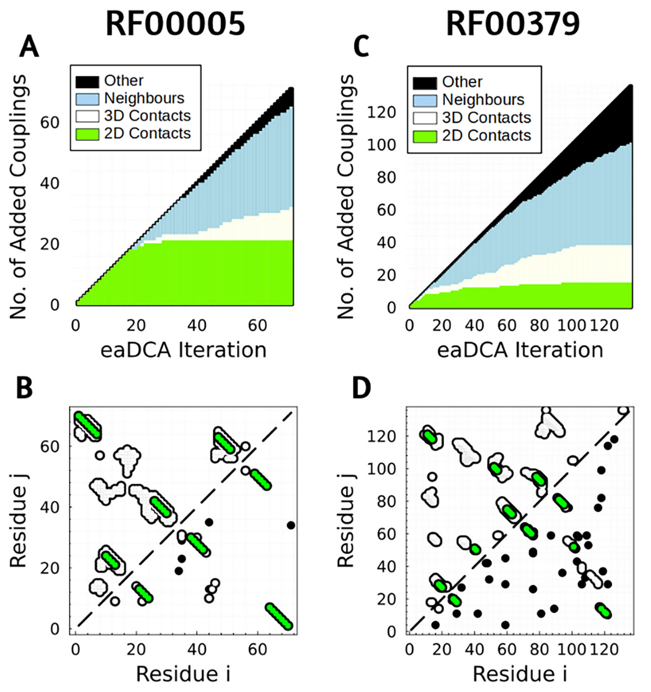


Figure 4. (**A**) RF0005: first L activated edges colored according to their classification. (**B**) RF0005: contact map (upper-left) and activated edges (lower-right). Secondary-structure contacts are evidenced in green, non-contacting activated edges in black. (**C**) RF00379: first L activated edges colored according to their classification. (**D**) RF00379: contact map (upper-left) and activated edges (lower-right). Secondary-structure contacts are evidenced in green, non-contacting activated edges in black.

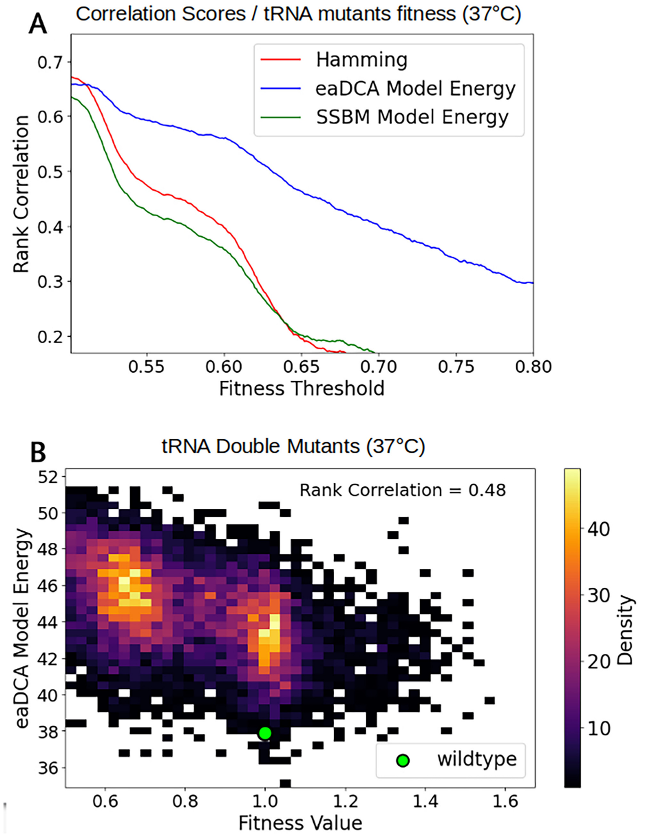


Figure 5. (**A**) Correlation of Hamming distance, eaDCA model energy and CM energy with tRNA fitness (37°C) at different values of minimum fitness threshold f_θ. (**B**) Relation between eaDCA model energy and tRNA fitness for the 8101 double mutants. For results at 23°C, 30°C cf. Supplementary Figure S7.

likelihood strategy used in their training assumes that nicely functional sequences have high probability, or equivalently low energy. Conversely, low-probability/high-energy sequences do not obey the evolutionary constraints learned by the model, and are expected to be non-functional.

This property can be used to predict mutational effects (12,13) by comparing the energies of the mutated and the wildtype sequences. In this way, a mutant sequence can be characterized by the energy difference

$$\Delta E = E(\text{mutant}) - E(\text{wildtype}).$$

A positive ΔE implies a reduction in the model probability for the mutant, suggesting that the mutation is likely to be deleterious. On the contrary, a negative ΔE signals a potentially beneficial mutation.

To test the quality of these predictions, we use the tRNA fitness dataset (27) (for details cf. *Materials and Methods*,Supplementary Section S5 and Supplementary Figure S6). We perform the following steps:

- For all mutant sequences in this dataset, we determine the energy differences to wildtype using both the eaDCA model, ΔE_{eaDCA}, and the SSBM model, ΔE_{SSBM}, as well as the Hamming distances (i.e. the number of mutations from wildtype).
- We select all mutant sequences having experimental fitness values $f \geq f_\theta$ above an arbitrary fitness threshold f_θ. This threshold is varied in our analyses to focus on diverse strengths of mutational effects.
- We calculate the Spearman rank correlation between the three predictors (eaDCA, SSBM, Hamming) and the fitness values f over the selected mutants, as functions of the fitness threshold f_θ.

As is shown in Figure 5A, when all mutants are included ($f_\theta = 0.5$), all three predictors show similarly good correlation values between 0.6 and 0.7. This results from the fact that most higher-order mutants, i.e. those of higher Hamming distance, have very low fitness, while mutants with one or two mutations frequently show more moderate fitness values. However, when increasing the fitness threshold f_θ, i.e. when including only mutations of more moderate fitness effects, ΔE_{eaDCA} correlations remain much more robust while the other two rapidly decay with f_θ. This shows that the eaDCA energies are informative over variable ranges of fitness effects.

To corroborate this finding, Figure 5B shows a heatmap of the 8101 two-point mutant sequences (at fixed Hamming distance of 2), comparing ΔE_{eaDCA} predictions and fitness values f. We observe a robust correlation even in this case, where the Hamming distance is constant and thus does not provide any information about the fitness measures.

Size estimation and constraint analysis of RNA sequence space

The entire space of sequences of given length is enormous. To illustrate this, the number of all ungapped sequences of length $L = 150$ is $4^{150} \simeq 2 \times 10^{90}$, if we include gaps like in our MSA, the number even rises to $5^{150} \simeq 7 \times 10^{104}$, and this exceeds by 10-24 orders of magnitude the estimated number $\sim 10^{80}$ of atoms in the universe. However, the viable sequence space related to a specific RNA family, i.e. to all sequences taking similar structure and performing similar function, is expected to be much smaller: sequences have to meet constraints imposed by residue conservation and coevolution, and possibly by other evolutionary constraints.

Our models allow for analyzing the impact of the different constraints on the entropy S and the size $\Omega = e^S$ of the sequence space, using the approach discussed in *Materials and Methods*. More specifically, the influence of conservation is measured via the entropy S_0 of the initial profile model, while the combined influence of conservation and coevolution is measured via the entropy S_{t_f} of the final model at termination (36). These results are corroborated by an independent estimation using a code published in (37), which estimates the size of the sequences space compatible with a given secondary structure, by efficiently sampling the neutral network related to a given RNA secondary structure.

The results are shown in Figure 6A for our selected RNA families. All three constraints enforce an exponential relationship between the size of the sequence space Ω and the sequence length L, i.e. the per-site reduction of the sequence space due to any individual type of constraint is roughly constant across the tested RNA families. Interestingly, conservation and secondary structure constrain the sequence space similarly, while the constraints imposed by both conservation and coevolution are, in line with expectations, the most stringent ones. As is illustrated in Figure 6B, out of the initially 5^L possibly gapped sequences of aligned length L about $(2.98 \pm 0.10)^L$ are compatible with the empirical conservation statistics, $(2.66 \pm 0.09)^L$ with the consensus secondary structure of the RNA families, and finally $(1.74 \pm 0.09)^L$ with both conservation and coevolution. To go back to our initial example $L = 150$, the final eaDCA sequence space would contain about 10^{36} distinct sequences: this number, while remaining enormous as compared to the observed extant sequences found in sequence databases like Rfam, comprises only a tiny fraction of 10^{-68} of the entire sequence space of this length, illustrating the fundamental importance of such constraints in the natural evolution of RNA families.

Note that these numbers also have an interesting interpretation in terms of the effective number of nucleotides, which are, on average, acceptable in a typical position of a functional RNA molecule. Out of the 5 theoretical possibilities (4 nucleotides or a gap), close to three are compatible with familywide conservation, or 2.66 with the consensus secondary structure. However, both constraints are insufficient for generative modeling as shown before. Our generative modeling indicates a much stronger reduction of the effective number of acceptable nucleotides per site to only 1.74 on average.

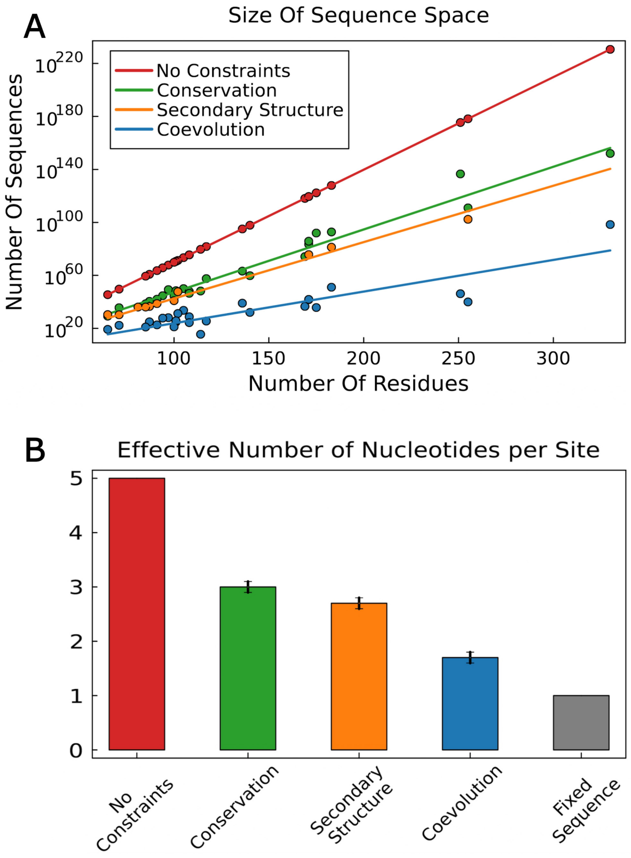


Figure 6. (**A**) Relations between RNA family length and the size of the sequence space under coevolution, conservation and secondary structure constraints. Dots are results for the different RNA families studied in this work, and lines are exponential fits. (**B**) Effective number x of nucleotides per site for each constraint. The size of the compatible sequence space is $\Omega = x^L$.

Structural characterization of artificial tRNA molecules by SHAPE-MaP probing

The definite test for the generative capacity of a statistical model of biomolecular sequences would involve expensive functional probing experiments on artificially sampled sequences. While this goes far beyond the scopes of our predominantly algorithmic paper, we have performed simpler and more cost efficient SHAPE-MaP experiments. SHAPE-MaP does not asses the functionality of the sequences, but provides non-trivial structural information: chemical probing reveals different reactivities for nucleotide positions, which are paired vs. unpaired in the secondary structure of the tested RNA molecule (31). SHAPE-MaP therefore allows us to check if our artificially generated sequences are compatible with the consensus secondary structure of the modeled RNA family, thereby corroborating the statistical tests described above.

The structural information obtained by such experiments is statistical: in the Reference dataset of published SHAPE experiments (cf. *Materials and Methods*), out of the paired sites typically >80% have low, less than 10% high

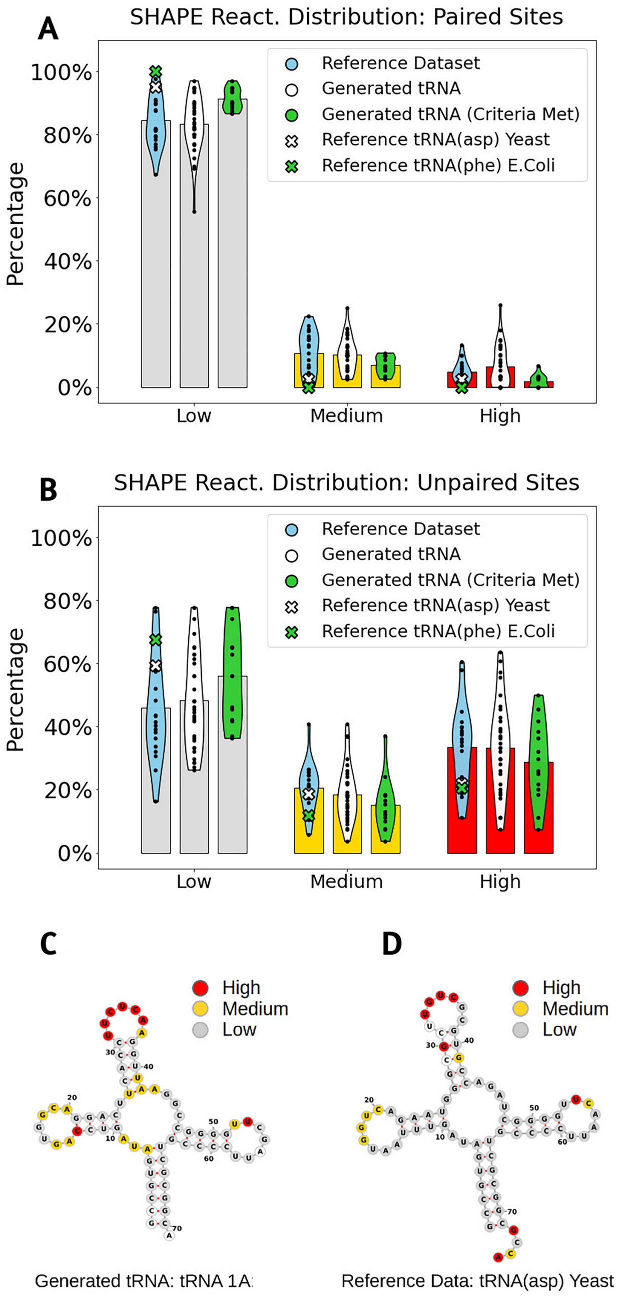


Figure 7. (**A**) Reactivity distribution for non paired residues, the bars refers to the indicated set average ('Reference SHAPE Dataset' $N = 21$, 'Generated tRNA' $N = 34$, 'Generated tRNA Criteria Me)' $N = 14$).(**B**) Reactivity distribution for paired residues. (**C**) Example of reactivity-structure projection for the 1A molecule of the 14 'Generated tRNA (Criteria Met)'. (**D**) Example of reactivity-structure projection for 'Reference SHAPE Data' tRNA(asp) Yeast

reactivity, while in the unpaired sites <50% have low and around 40% have high reactivity, cf. Figure 7. Determining the specific pairing status of individual pairs is non-trivial due to a number of confounding factors: first, the correlation between SHAPE reactivity and base pairing is nonlinear. Second, SHAPE data may not mirror a single structure, but an average reactivity across a structural ensemble. Third, SHAPE reactivity can also be influenced by factors beyond secondary structure, such as base stacking and tertiary contacts (31). Nevertheless, SHAPE-MaP experiments are a valuable instrument for assessing whether the SHAPE profile of a tested RNA molecule—natural or artificial—is statistically coherent with its expected secondary structures.

We used the tRNA family (RF0005) discussed above for generating 76 artificial tRNA sequences. We probed the SHAPE reactivities at each site of these sequences (for in depth details about the experiment refer to Supplementary Section S6 and Supplementary Table S5). We categorized the reactivities into three classes: low, medium, and high (as is common practice (31,33)) and we assessed the distribution of these classes among paired and unpaired residues for each sequence. Due to experimental reasons we did not sample the 76 sequences freely from $P(a_1, ..., a_L)$, but we introduced two types of constraints:

- Due to experimental constraints, the last 16 nucleotides were kept constant, cf. *Materials and Methods* and Supplementary Section S6. Only the first 55 positions were generated by the model conditioned to the last 16, i.e. they were sampled from $P(a_1, ..., a_{55} | a_{56}, ..., a_{71})$. This reduces the effective sequence space $\Omega = e^S$ from $\sim 10^{22}$ sequences (cf. Table 1) to $\sim 10^{14}$, which is still a huge number beyond the possibility of exhaustive testing.
- Inspired by works about proteins (14,38) and aiming at increasing the success rate in a limited number of experiments, only sequences of low energy ($E < 44$) and good secondary-structure score ($F > 0.53$, measured as the F-score between the RNAfold (October 2022) (39) predicted structure and the tRNA consensus one) were included in the test, cf. the details given in the Supplementary Section 6.1. These filters come at relatively low cost: while the energy-based filter is met by about 50% of all sampled sequences, the double filter still preserves about 20% of the sequences (Supplementary Figure S8), inducing thus a very moderate decrease of the size of the sequence space.

For a more detailed overview of the dataset used in the test, please refer to the Supplementary Section S6.2 and Supplementary Figures S8, S9, S10. Finally after probing, 34 of the 76 tRNAs satisfied the experimental standard of possessing reactivity data for more than 50% of the sequence positions and were included in our further analyses.

In Figure 7, we observe that the reference dataset and the generated tRNA behave similarly, with clearly visible differences between paired and unpaired sites. We employed Permutational Multivariate Analysis of Variance (PERMANOVA) to test for statistical differences between the reactivity distributions of the reference dataset and of the generated tRNA, and between paired and unpaired sites. While we do not see indications for statistically significant differences between the reference dataset and the generated sequences (P-values of 0.993 for paired sites, 0.420 for unpaired sites), the paired and unpaired sites in the generated sequences are significantly distinct (P-value 1.9×10^{-7}).

Moreover, observing that the statistics for paired residues are more rigorous, especially on the two 'Reference SHAPE dataset' tRNA, we decided to implement an additional filtering criterion. We deem artificial molecules as 'Criteria Met' if over 85% of their paired residues fell into the low reactivity class. 14 out of 34 generated tRNA are classified as 'Criteria Met'. Those are also the sequences that better pass the qualitative visual criterion (Figure 7C, Supplementary Section S6.3.4 and Supplementary Figure S11).

These results, albeit rather qualitative, indicate that the SHAPE reactivities of our artificially generated tRNAs are as consistent with the desired tRNA secondary structure, as the sequences in the reference dataset are with their published secondary structures. While these results generally support the eaDCA model as a valid generative model, a more comprehensive experimental evaluation might include (i) control groups generated by simpler models (e.g. profile, SSBM, CM) to assess if eaDCA is necessary to generate structurally coherent sequences, (ii) a filter-free generation directly using the statistical models, to test the model in an unbiased way and (iii) functional rather than structural tests to fully assess the model's generative capacities.

Conclusion and outlook

As in many disciplines and thanks to the strong increase in data availability, generative models gain growing importance also in the modeling of biomolecular sequences. A first practical reason is quite obvious: generative models are of high biotechnological interest in biomolecular optimization and *de novo* design, directly or in combination with screening or selection assays when suggesting functionally enriched sequence libraries.

A second reason is less obvious, but has the potential to be at the basis of a paradigmatic shift in computational molecular biology. Traditionally, sequence bioinformatics was dominated by simpler statistical models, like the profile or covariance models discussed also in this paper, and which are of great success in analyzing extant biomolecular sequences, detecting homology, annotating sequences functionally, establishing RNA or protein families, reconstructing their phylogenies or aligning sequences. Generative models have the potential to go substantially beyond this, and to directly contribute to our future understanding of biological molecules in their full complexity as high-dimensional, disordered and interacting systems. When a model is capable to generate diversified but viable artificial sequences, it necessarily incorporates essential constraints, which are functionally or structurally imposed on the sequences in the course of evolution. Even in this case, there is no guarantee that only such essential constraints are present in the model, and that these are encoded in a biologically interpretable way. In our work, we are therefore searching for *parsimonious* models, which contain as few as possible useless constraints (by using an information-theoretic criterion for including constraints, or the corresponding parameters, into the modeling), and which in turn should be maximally interpretable.

However, generative modeling is not trivial. The total sequence space is enormous, while the example sequences in RNA or protein family databases are quite limited. Very different models may be generative. In a parallel effort, (9) proposed and experimentally validated restricted Boltzmann machines (RBM) as generative models. In the case of protein families, it was shown before that RBM, which are shallow latent-space models, are able to detect extended functional sequence motifs (40,41), but at the same time they have difficulties in representing pairwise structural constraints like residue contacts. On the contrary, eaDCA was found to easily detect contacts, but the patterns responsible for the clustered structure of families into subfamilies, easily visible by dimensional-reduction techniques like principal component analysis, remain hidden in the coupling network. It remains a challenge for the future to combine such different approaches to further improve interpretability of generative models.

Another problem is that, by definition, generative models reproduce statistical features of the training data, but there is no guarantee that statistical similarity implies biological functionality - this dilemma is well known from text or image generation with generative models, which do not always produce correct text contents or possible images, and extensive experimental testing is needed to fully prove the validity of the models.

Finally, an intrinsic limitation of data-driven sequence models is that, even if introducing substantial novelty into generated data when compared to training data, they are unable to extrapolate to regions of sequence space that are are *a priori* viable, but not reached by natural evolution, or not present in sequence databases. To explore such regions, it may be necessary to go beyond single Rfam families to unveil generic evolutionary constraints acting across families, or to include non-data-driven constraints (e.g. biophysical folding models) into the generative models.

Despite such limitations, generative modeling will naturally benefit from the current evolution of more and more quantitative high-throughput experimental approaches in biology. On one hand, these can be used naturally to test model predictions (e.g. mutational effects) and sequences generated by the models, going far beyond the low-throughput experiments we presented in this predominantly computational work. On the other hand, these techniques substantially change the data situation in biology in several aspects (cf. e.g. (14,27)): while current dataset, i.e. MSA of homologous RNA or protein families, consist of positive but experimentally non annotated data, experiments provide (i) quantitative functional annotations for thousands of sequences and (ii) negative examples for artificial non-functional sequences generated by imperfect methods like random mutagenesis or sampling from imperfect models learned from finite data. This change in data will trigger future methodological work to develop integrative methods using all biologically relevant available information within the modeling process.

Data availability

The data and the version of the code used at the time of this study are available at DOI: 10.5281/zenodo.10688226.

For any subsequent updates refer to this Github repository: https://github.com/FrancescoCalvanese/FCSeqTools.jl.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We are grateful to Sabrina Cotogno, Matteo Bisardi, Roberto Netti and Vaitea Opuu for helpful discussions during the project and the writing of the paper. We acknowledge also funding by the Institut Pierre-Gilles de Gennes (ANR-10-EQPX-34, to P.N.), EU H2020 Grant ERC AbioEvo (101002075, to P.N.), Human Frontier Science Program (RGY0077/2019, to P.N.), EU H2020 grant MSCA-RISE InferNet (734439, to M.W.). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945304 – Cofund AI4theSciences hosted by PSL University.

High-throughput sequencing was performed by the ICGex NGS platform of the Institut Curie supported by the grants ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) from the Agence Nationale de la Recherche ('Investissements d'Avenir' program), by the ITMO-Cancer Aviesan (Plan Cancer III) and by the SiRIC-Curie program (SiRIC Grant INCa-DGOS-465 and INCa-DGOS-Inserm-12554). Data management, quality control and primary analysis were performed by the Bioinformatics platform of the Institut Curie.

Funding

H2020 European Research Council [AbioEvo/101002075]; H2020 Marie Sklodowska-Curie Actions [InferNet/734439]; Agence Nationale de la Recherche [ANR-10-EQPX-34]; Human Frontier Science Program [RGY0077/2019]; H2020 Marie Sklodowska-Curie Actions [AI4theSciences/945304].

Conflict of interest statement

None declared.

References

1. Holoch,D. and Moazed,D. (2015) RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.*, **16**, 71–84.
2. Castel,S.E. and Martienssen,R.A. (2013) RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.*, **14**, 100–112.
3. Walter,N. and Engelke,D. (2002) Ribozymes: Catalytic RNAs that cut things, make things, and do odd and useful jobs. *Biologist*, **49**, 199–203.
4. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z., *et al.* (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, **49**, D192–D200.
5. Leonardis,E., Lutz,B., Ratz,S., Cocco,S., Monasson,R., Schug,A. and Weigt,M. (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, **43**, 10444–10455.
6. Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C. and Marks,D.S. (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.
7. Pucci,F., Zerihun,M.B., Peter,E.K. and Schug,A. (2020) Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. *RNA*, **26**, 794–802.
8. Cuturello,F., Tiana,G. and Bussi,G. (2020) Assessing the accuracy of direct-coupling analysis for RNA contact prediction. *RNA*, **26**, 637–647.
9. Fernandez-de-Cossio-Diaz,J., Hardouin,P., du Moutier,F.-X.L., Gioacchino,A.D., Marchand,B., Ponty,Y., Sargueil,B., Monasson,R. and Cocco,S. (2023) Designing molecular RNA switches with restricted Boltzmann machines. bioRxiv doi: https://doi.org/10.1101/2023.05.10.540155, 12 May 2023, preprint: not peer reviewed.
10. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
11. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
12. Figliuzzi,M., Jacquier,H., Schug,A., Tenaillon,O. and Weigt,M. (2015) Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.*, **33**, 268–280.
13. Levy,R.M., Haldane,A. and Flynn,W.F. (2017) Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.*, **43**, 55–62.
14. Russ,W.P., Figliuzzi,M., Stocker,C., Barrat-Charlaix,P., Socolich,M., Kast,P., Hilvert,D., Monasson,R., Cocco,S., Weigt,M., *et al.* (2020) An evolution-based model for designing chorismate mutase enzymes. *Science*, **369**, 440–445.
15. Schneider,B., Sweeney,B.A., Bateman,A., Cerny,J., Zok,T. and Szachniuk,M. (2023) When will RNA get its AlphaFold moment?. *Nucleic Acids Res.*, **51**, 9522–9532.
16. Figliuzzi,M., Barrat-Charlaix,P. and Weigt,M. (2018) How pairwise coevolutionary models capture the collective residue variability in proteins?. *Mol. Biol. Evol.*, **35**, 1018–1027.
17. Muntoni,A.P., Pagnani,A., Weigt,M. and Zamponi,F. (2021) adabmDCA: adaptive Boltzmann machine learning for biological sequences. *BMC Bioinformatics*, **22**, 528.
18. De Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
19. Cocco,S., Feinauer,C., Figliuzzi,M., Monasson,R. and Weigt,M. (2018) Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.*, **81**, 032601.
20. de la Paz,J.A., Nartey,C.M., Yuvaraj,M. and Morcos,F. (2020) Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 5873–5882.
21. Bisardi,M., Rodriguez-Rivas,J., Zamponi,F. and Weigt,M. (2021) Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *Mol. Biol. Evol.*, **39**, msab321.
22. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E1301.
23. Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.
24. Madani,A., Krause,B., Greene,E.R., Subramanian,S., Mohr,B.P., Holton,J.M., Olmos,J.L., Xiong,C., Sun,Z.Z., Socher,R., *et al.* (2023) Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, **41**, 1099–1106.
25. Barrat-Charlaix,P., Muntoni,A.P., Shimagaki,K., Weigt,M. and Zamponi,F. (2021) Sparse generative modeling via parameter reduction of Boltzmann machines: application to protein-sequence families. *Phys. Rev. E*, **104**, 024407.
26. Zerihun,M.B., Pucci,F. and Schug,A. (2021) CoCoNet—boosting RNA contact prediction by convolutional neural networks. *Nucleic Acids Res.*, **49**, 12661–12672.
27. Li,C. and Zhang,J. (2018) Multi-environment fitness landscapes of a tRNA gene. *Nat. Ecol. Evol.*, **2**, 1025–1032.
28. Hajdin,C., Bellaousov,S., Huggins,W., Leonard,C., Mathews,D. and Weeks,K. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5498–5503.
29. Cocco,S. and Monasson,R. (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys. Rev. lett.*, **106**, 090601.
30. Barton,J.P., De Leonardis,E., Coucke,A. and Cocco,S. (2016) ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, **32**, 3089–3097.
31. Kutchko,K.M. and Laederach,A. (2016) Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *WIREs RNA*, **8**, e1374.
32. Busan,S. and Weeks,K. (2017) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA*, **24**, 143–148.

33. Bellaousov,S., Reuter,J.S., Seetin,M.G. and Mathews,D.H. (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.*, **41**, W471–W474.
34. Trinquier,J., Uguzzoni,G., Pagnani,A., Zamponi,F. and Weigt,M. (2021) Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.*, **12**, 5800.
35. Rivas,E., Clements,J. and Eddy,S.R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.
36. Barton,J.P., Chakraborty,A.K., Cocco,S., Jacquin,H. and Monasson,R. (2016) On the entropy of protein families. *J. Stat. Phys.*, **162**, 1267–1293.
37. Jörg,T., Martin,O. and Wagner,A. (2008) Neutral network sizes of biological RNA molecules can be computed and are atypically large. *BMC Bioinformatics*, **9**, 464.
38. Malbranke,C., Bikard,D., Cocco,S. and Monasson,R. (2021) Improving sequence-based modeling of protein families using secondary-structure quality assessment. *Bioinformatics*, **37**, 4083–4090.
39. Lorenz,R., Bernhart,S.H., zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithm. Mol. Biol.*, **6**, 26.
40. Tubiana,J., Cocco,S. and Monasson,R. (2019) Learning protein constitutive motifs from sequence data. *eLife*, **8**, e39397.
41. Shimagaki,K. and Weigt,M. (2019) Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Phys. Rev. E*, **100**, 032128.

Received: October 14, 2023. **Revised:** March 5, 2024. **Editorial Decision:** March 31, 2024. **Accepted:** April 5, 2024