# Hybrid Generative Model: Bridging Machine Learning and Biophysics to Expand RNA Functional Diversity

Vaitea OPUU[a,b,‡]

[a]Laboratoire de Biophysique et Evolution, UMR CNRS-ESPCI 8231 Chimie Biologie Innovation, PSL University, France,
[b]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[1]vaitea.opuu@espci.fr

**Abstract**

The RNA world hypothesis suggests that RNA once catalyzed reactions now performed by proteins. Rediscovering these functions requires exploring sequence spaces beyond natural RNAs. While machine learning (ML)-based RNA design shows promise, it struggles to extrapolate beyond training data. In contrast, biophysics-based approaches leveraging RNA secondary (2D) structure operate independently of training data but are not tailored for functional discovery. We present a hybrid generative model that combines a Potts model with the thermodynamic folding model of RNA 2D structure. This approach disentangles folding contributions from functional signals, such as binding, enabling the data-driven component to focus on tertiary interactions and improving contact predictions. This disentanglement introduces structural imprinting—a novel strategy that uses structural variability to guide mutations, which showed great promise in uncovering hidden natural diversity. By bridging ML and biophysics, this model tackles the longstanding challenge of expanding diversity beyond the mere reproduction of the training data.

## Introduction

RNA is a central biomolecule in biology, performing dual roles as a carrier of genetic information and a biocatalyst in critical processes such as ribosomal catalysis, splicing, and gene regulation [10]. These capabilities support the RNA world hypothesis, which proposes that early life relied on RNA to perform catalytic functions now undertaken by proteins [6]. This hypothesis implies that RNA has untapped functional potential, with many stable folds and sequence spaces remaining unexplored [9]. Advances in RNA technologies have demonstrated the practical significance of this potential. Applications such as biosensors [13], gene modulation [30], and RNA-based therapies, including CRISPR systems and mRNA vaccines [26], illustrate the transformative impact of RNA in biotechnology and medicine. Furthermore, in vitro selection experiments have revealed the ability of RNA to perform catalytic activities not observed in nature [10], emphasizing the vast possibilities for RNA engineering. A major challenge in advancing RNA research is understanding sequence-structure-function relationships. Exploring RNA sequence spaces beyond those found in nature requires computational tools that enable the rational design of functional RNAs, facilitating the discovery of novel activities and therapeutic applications.

Designing catalytic RNAs is particularly challenging compared to proteins. Machine Learning (ML) has achieved near-experimental accuracy for protein structure prediction [12], enabled by a vast dataset of protein 3D structures and the absence of new structural motifs, making sequence-structure prediction well-posed. This success has led to generic algorithms capable of designing proteins for diverse natural functions [8, 29]. In contrast, RNA data is significantly scarcer and biased, with about 36 million sequences in the RNACentral database [1] and a few thousands structures. Consequently, ML applications to RNA remain limited, typically restricted to introducing mutations mimicking natural family variability [3, 25].

Physics-based approaches are widely used to study phenomena lacking direct measurements, as they primarily rely on first principles. However, these methods are often computationally expensive and fail to account for critical biological constraints. For example, molecular dynamics (MD) simulations, which model atomic

motions, are limited to short timescales (below microseconds), whereas biological processes like folding typically occur over milliseconds. To address these limitations, integrating evolutionary information has proven effective. Schug et al. [24] demonstrated this by combining genomic contact residue data with molecular simulations, achieving high-resolution protein complex models. Despite their utility, physics-based approaches have shown lower success rates in protein design compared to machine learning methods, with activity rates as low as $\sim$ 6% in some cases [4]. Nevertheless, strategies such as unfolded state models parameterized with evolutionary diversity have successfully redesigned PDZ domains [28], generating sequences with native-like properties. Furthermore, recent advancements have expanded enzyme functionality into unexplored regions. For instance, methionyl-tRNA synthetase was redesigned to incorporate $\beta$-methionine, a noncanonical amino acid [27], highlighting the potential of physics-based approaches in data-scarce scenarios.

RNA design and prediction are greatly facilitated by base-pairing interactions, which enable algorithms to predict secondary (2D) structures by minimizing folding free energies derived from physical principles [16]. These methods efficiently identify structural elements, such as stems and loops, though accurate prediction of three-dimensional structures remains a significant challenge [32]. To address data scarcity, models incorporating biophysical a priori principles have been proposed. For instance, MxFold integrates thermodynamic principles to enhance RNA 2D structure predictions [23]. In RNA design, hybrid models that combine machine learning with biophysical principles have expanded the exploration of self-reproducing RNA sequence spaces beyond the diversity observed in training data [15]. This approach adds the biophysical folding contribution *a posteriori* to a shallow ML model trained on natural sequences, enabling the sampling of novel designs. While promising, these efforts underscore the need for hybrid models with automatic training methods that explicitly disentangle biophysical contributions from data-driven components to further advance the design of functional RNAs.

Here, we propose a hybrid RNA design framework combining an ML model (Potts model) with the thermodynamic folding model of RNA 2D structures. Incorporating the folding model into the ML training process allows for the disentanglement of folding contributions from the data-driven component. By explicitly introducing folding, the ML (data-driven) component focuses on capturing non-trivial signals, typically associated with tertiary (3D) interactions and function.

We first demonstrate that our approach improves the inference of non-trivial nucleotide interactions from multiple sequence alignments. Using simulated data, we then show that the hybrid framework effectively separates biophysical folding contributions from functional signals, enabling the design of RNAs specifically optimized for binding. This includes the ability to imprint binding functionality onto alternative structural configurations. Finally, leveraging these properties, we show that our approach can steer the sampling of functional RNA sequences beyond the training data, using natural RNA sequences as the basis for this exploration.

## The hybrid model

The distribution of functional RNA is tightly bound to folding thermodynamics because, as with proteins, the structure shapes binding and catalytic sites. However, ML models for RNA sequences do not account explicitly for folding thermodynamics. Here we propose a hybrid ML model that has i) a biophysical component explicitly accounting for folding and ii) a data-driven component for fitting the input MSA to what is not explained by the biophysical component. Here, we implement this approach by combining the Direct Coupling Analysis (DCA) ML model with the nearest neighbor folding model.

The DCA generative model was initially proposed for the inference of protein contact maps [19]. DCA takes as input a multiple sequence alignment (MSA). It is based on a Potts model that approximates the joint probability distribution $P(\mathbf{S})$ over sequences $\mathbf{S} = (S_1, S_2, \ldots, S_L)$, where $L$ is the sequence length, and $S_i$ represents the nucleotide at position $i$. The distribution is expressed as:

$$\begin{aligned} P(\mathbf{S}) &= \frac{1}{Z} \exp\left(\frac{1}{T} H(S)\right) \\ H(S) &= \sum_{i<j} J_{ij}(S_i, S_j) + \sum_i h_i(S_i), \end{aligned} \tag{1}$$

where $J_{ij}(S_i, S_j)$ are the coupling parameters representing direct interactions between positions $i$ and $j$,

$h_i(S_i)$ are the site-specific terms controlling propensities, and $Z$ is the partition function ensuring normalization. $T$ is the sampling temperature. The model is trained by maximizing the likelihood of the sequences observed in the MSA using, that can be done with various training algorithms [5]. Once the parameters are trained, the sequences are sampled from the model using Markov Chain Monte Carlo. For contacts inference $S(i,j) = \sqrt{\sum_{a,b} J_{ij}(a,b)^2}$ we compute the Frobenius norm of the couplings of all 25 pairs of nucleotides $a, b$ at positions $(i, j)$ (including the gap). Then, we apply the Average Product Correction (APC) to obtain the coupling score [5], where we choose the $k$ highest $S(i,j)$ for contact prediction. In our results, we expressed the number of ranked scores as a fraction of the length $L$ of the MSA.

The biophysical model of RNA 2D structure, called the nearest neighbor, allows calculating its thermodynamic stability. It assumes that the folding free energy $\Delta G(S|c)$ of a sequence $S$ from the unfolded state to the folded conformation $c$ depends on the stacking interactions in stems and penalties from unpaired nucleotides in loops (bulge, internal loops, and hairpins) [18]. The biophysical model uses experimentally determined parameters to compute $\Delta G(S|c)$ for all possible configurations. Using this energy function, the minimum free energy (MFE) structure of a given sequence is identified by dynamic programming [33]. In an MSA, a consensus conformation can be extracted using the approach implemented in the RNAalifod [2], which accounts for stability but also sequence conservation.

In this work, we combine the DCA with the folding energy model:

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\frac{1}{T}(H(S) + \phi \Delta G(S|c))\right), \tag{2}$$

where $\phi$ is a scaling parameter that disentangles the structural contribution of the folding $\Delta G(S|c)$ into the configuration $c$ from the data-driven contribution that is DCA $H(S)$, see Fig 1a. The configuration $c$ is an input of the model which can be for example the consensus structure of the MSA or the MFE of one of the sequence. The parameters of $H(S)$ are trained using the Boltzmann learning algorithm with the $L_1$ regularization [7]. The scaling parameter $\phi$ is optimized simultaneously following its gradient with respect to the sequence log-likelihood $L = \frac{1}{D}\sum_{d=1}^{D} \log P(S^d)$:

$$\frac{\partial L}{\partial \phi} = \langle \Delta G(\mathbf{S}) \rangle_{MSA} - \langle \Delta G(S) \rangle_{P(S,\phi)}. \tag{3}$$

The first term is the average folding free energy across the training sequences whereas the second term is the expected folding free energy averaged with respect to the probability given by the model. Similarly to [7], in the contrastive version of the Boltzmann learning, we use the $\Delta G(S|c)$ of the sequence sampled at time $t$ as an instantaneous estimate of the average of the model. Now that the model is defined as the sum of the data-driven and biophysical contribution, we can assemble the trained DCA with a different configuration $c*$ to steer the sampling of sequences in regions where these new structural constraints are satisfied Fig 1b. We show in Fig 1c an example of the DCA part extracted from a hybrid model trained on the tRNA MSA from RFAM, where the 2D interactions are attenuated.

An important caveat of our approach is that RNA MSAs contain gaps which do not participate in the folding energy; therefore, in trimmed MSAs to only extract the positions for a target sequence, the folding energy is an heuristic contribution because the non-aligned positions in the homologous sequences are not contributing to the folding. However, as shown in later in our experiments, this heuristic still carries enough structural informations.

The proposed model is designed to disentangle structural stability from the remaining signal, which is thought to be related to the activity. By achieving this, our approach enables the independent assembly of data-driven and structural contributions. This unique property introduces an unprecedented capability: varying structural conformations to explore the sequence space more effectively.

## Non trivial contacts predictions

RNA 2D structure contacts are trivial to predict with DCA from MSAs because educated guesses of the 2D structure are often explicitly used in the construction of the MSA, *e.g.* INFERNAL [20]. In contrast, non-trivial

contacts, such as weaker 3D interactions, remain challenging to predict. Here, introducing the 2D structure stability in the DCA should allow the model to focus on non-trivial contacts, which are eventually related to the function such as binding. Ultimately, this approach will improve the generation of molecules by capturing non trivial signal that remain undetected by other approaches.

We test now that our hybrid DCA approach is able to improve the non-trivial contact inference by integrating the 2D structure stability explicitly. To do so, we started with the curated dataset of RFAM [14] families of homologous natural RNAs presented in [22]. We extracted the 57 families following [31], with sequence length ranging from 41 to 496 nucleotides. To each RFAM family is associated a PDB 3D structure allowing the determination of the experimental 2D structure using DSSR [17]. The MSAs were trimmed such that only the positions of the 3D structure were kept, which we refer to as the target sequence. In the 3D structure, two nucleotides are considered in contact if one of their heavy atoms are 10 $\mathring{A}$ apart. From this, we defined non-trivial contacts as all the contacts that are not in the 2D structure, including pseudoknots.

We trained one instance of our hybrid model and one of the regular DCA model (with $\phi = 0$) for each of the 57 families selected. For the stability contribution of the hybrid model, we considered three types of structural configurations: EXP) the experimental 2D structure, MFE) the predicted minimum free energy structure of the target sequence, and ALI) the consensus structure obtained from the MSA using the tool RNAalifold. For contact inference, we set the pairs of positions involved in the 2D configurations used (MFE, ALI, or EXP) with the highest coupling scores so that they are necessarily added to the predicted contacts. Then, we inferred contacts using the couplings, where we extract the top coupling scores across varying fraction of $L$, that is the length of the sequence. The performance is measured by the positive predicted value (PPV), which is the fraction of correctly predicted contacts.

Our null model, the DCA not informed by folding, is shown in blue in Fig. 2. It predicts all contacts types effectively, with an average PPV ranging from 0.8 to 0.39 as the tested fraction increases from 0.1 to 1 $L$ (where we capped at 1 $L$ since the maximum number of 2D contacts is inherently limited to $L/2$). This results mean that the top 10\of true contacts. For only 2D contacts, the accuracy range from 0.68 to 0.15. These 2D predictions are actually lower than the fully biophysics-based prediction obtained with RNAfold [16] (Fig 5), which demonstrate that these contacts are alerady quite well predicted by non-ML algorithms. Then, we benchmarked the non trivial 3D contacts. To do so, for each prediction, we first remove the pairs of positions that are in the EXP configuration by setting them to zero in the predicted contact map. The PPV obtained for non-trivial contacts is drastically lower, averaging between 0.51 and 0.28 across the same range of fractions of $L$. A critical factor contributing to successful contact prediction is regularization, which, in our implementation, influences tertiary predictions, particularly beyond 0.5 $L$, as shown in Fig. SIX.

Our hybrid DCA model with the ALI configuration improved the contact prediction significantly in every prediction types (ALL, 2D, and 3D) compared with the regular DCA, as demonstrated by one-sided Wilcoxon tests (pvalues $<$ 0.05 in Fig 2). The ALI configuration is considered as the most representative fold because it accounts for structural constraints of all the sequences in the MSA. To validate this, we analyzed the distribution of $\phi$. This parameter trained similarly to the data-driven part represents the importance of the structural stability in the hybrid model (the higher the more important the structure is). Fig. 2d shows that on average, the ALI configuration aligns better with the MSAs, supported a one-sided Student t-test with p-value $< 10^{-5}$. When the 2D structure does not match the MSA, $\phi$ approaches zero, indicating that the data-driven component of the hybrid model compensates by fully learning the signal from the MSA. Thus, the biophysical component should enhance the ability of the data-driven component to capture non-trivial structural signals if it aligns well with the MSA.

The MFE configuration represents the most stable structure predicted by the biophysical model for the target sequence. Applying our hybrid model with this configuration yielded higher ALL and 2D contact predictions compared to the ALI configuration but resulted in lower 3D predictions. This discrepancy arises because other sequences in the MSA may exhibit slightly different MFE configurations, leaving some 2D contacts in the MSA unaccounted for by the MFE of the target sequence alone. Consequently, the hybrid MFE model shifts its focus to the unaccounted 2D pairs. Similarly, the hybrid model using the experimental EXP configuration also failed to improve 3D predictions, as the EXP structure of the target sequence also does not contain all 2D contacts in the MSA. However, the EXP configuration achieved the highest performances in ALL and 2D predictions because it includes all the true 2D contacts. These findings confirm that the hybrid model requires a configuration that recapitulates the MSA and not the target sequence.

Finally, we evaluated the hybrid approach using a non-generative DCA variant called Mean Field DCA (mfDCA), which computes coupling parameters by inverting the covariance matrix of the MSA [19]. To integrate structural information, we reweighted sequence contributions to the covariance matrix with their folding stability $\Delta G(s|MFE)$ (described in SI). For each RNA family, we trained both the biophysically informed MfDCA and the regular mfDCA, but the biophysically informed variant did not improve structure prediction (Fig. 6). Additionally, we tested the Pseudolikelihood DCA (PLM) [11] variant, which marginally outperformed mfDCA (Fig. 5); however, we did not implement its hybrid version because we suspect that the reweighting scheme proposed is not introducing enough structural information, which highlights the necessity of using generative DCA variants.

Here, we showed that our hybrid model learns more from the same data by effectively lowering the focus off the trivial contacts that are already quite well already predicted with the biophysical folding model. One part of the signal captured is related to tertiary interactions, which are correlated to binding and catalysis in ribozymes. Therefore, this approach should recapitulate better function, which ultimately should improve the generation of functional RNAs.

## Stability and function disentanglement

Optimizing RNA design for specific functions, such as binding or catalysis, is a significant challenge because these contributions are not disentangled in training data; therefore, ML-based generative models often fail to design for specific properties. Here, we demonstrate that hybrid models can overcome this limitations by effectively disentangling molecular stability from binding, allowing the design to focus on optimizing function.

We simulated a dataset with a structured RNA of 29 nucleotides that binds to a four-nucleotide ligand that we constructed artificially, see configuration A in Fig 3. We performed *in silico* random mutagenesis and selection based on folding and binding. The selected mutated RNAs is the training data we use to demonstrate that our hybrid strategy is able to disentangle stability from function/binding, see the details of the artificial data construction in SI.

We trained our hybrid model on the artificial data produced. To account for the local nature of the training data, we explicitly introduce the diversity term $\omega$ described in section SI, acting as an effective regularization for the poor diversity. This term essentially assumes that introducing mutations is exponentially deleterious for the function, which is controlled by the parameter $\omega$. We show in Fig X the results of all tested design strategy for binding. The binding score measures how well the ligand binds to the 5' end of each sequence, see details in SI.

First, we show that even starting with a dataset with as few as maximum 10 mutations, we can generate binders with more up to 20 mutations with our hybrid model (HYB), which outperformed the regular DCA (REG), see Fig. 3. Typically, this type of gain is explained by a loss of diversity [15], where the model focuses on a group of close mutants that are active, which is done by reducing the sampling temperature $T$. To measure the diversity produced, we compute the maximum Hamming distance between two mutants at each distance to the WT, which we refer to as the design diameter in Fig 3. The sequence diversity generated by the hybrid model is similar to the one generated by the regular DCA, see Fig 3c. As expected, reducing the $T$ of the regular DCA (REG(T=0.6)) increased the active fraction; however, this increase is made at the cost of sequence diversity. The later shows that our model indeed generates a larger array of functional binders.

How to optimize the binding specifically? The above improved generation is thought to be the result of a successful stability versus binding disentanglement, where the binding is modeled by the data-driven term, which is DCA. Thus, by increasing the weight of the data driven term in the sampling of sequences, we should be able to generate a higher fraction of active binders. By doing so, we indeed design higher rates of binder of design while preserving the sequence diversity (Fig 3). This result confirm that our model disentangled stability from function.

Designing using only the biophysical property is shown in expected to be low because the function is rarely captured by the biophysical model. Here, we investigated this question by designing mutants using only the folding contribution (STRUCT). Our results show that while the structure of the designs is well preserved, unsurprisingly less binders are found. Therefore, our approach is a synergistic combination of biophysics and data-driven approaches.

If the two contributions are truly independent, we should be able to vary the structural configuration while maintaining the function. To test this, we used the hybrid model trained earlier on the configuration A and extracted the data-driven component. This component is assumed to have absorbed the binding specifically. Next, we combined the pre-trained data-driven component with the folding into the configuration B (Fig 3). This configuration is totally different from A (the 5' end is left unpaired) and for which we do not have a single training sequence. The goal is to steer the sampling toward sequences that now fold into configuration B but that are still binding the ligand. In Fig 3e, we show that we still generate active binders, but that now fold into configuration B. In Fig 3d, we show that these sequences interpolate the space between active binders in configuration A (the starting point) to active binders in configuration B. These results demonstrate that in an ideal case, we can imprint the function on compatible alternative configurations.

Our findings demonstrate that hybrid models combining biophysical folding and data-driven components effectively disentangle function from 2D stability, allowing us to optimize RNA specifically for function. Leveraging, this property, we introduced structural imprinting of function onto alternative configurations. This approach challenges the conventional protein/RNA design paradigm of "same structure, same function" by introducing a novel form of design diversity—structural diversity.

## Extrapolation by structural imprinting

The greatest challenge in ML is generalization. Generative models typically fall short on this task because they are trained to reproduce the diversity observed in the training data; therefore limiting the exploration of new sequences. We showed above that by introducing a biophysical component to our model, we effectively disentangled its contribution from the data-driven component. The result is an approach allowing us to enhances sequence and structure design diversity. Leveraging these findings, we demonstrate here that our approach helps explore the sequence space of functional natural RNAs, which represents a critical step in recovering the lost RNA functional and structural diversity.

To demonstrate this, we started with 25 MSAs of natural RNA families extracted from [22] with families of at least more than 200 sequences (but less than 5000 for the sake of computational resources) and length less than 300 nucleotides as we expect a drop in modeling accuracy [**?**]. Then, we cluster the sequences of each MSA into two clusters using the spectral clustering algorithm implemented in Sci-kit learn (version 1.0.2), based on the hamming distance between sequences. For each pair of clusters, we computed two consensus structures, which we refer to as configuration A and B. The average base pair distance between the two is $\sim$ 7, showing that the MSAs are typically composed of structurally consistent sequences. The goal of this is to test whether the model trained only on one of the sub MSAs is able to generate sequences from the second sub MSA. By doing so, we would demonstrate that it is possible to extrapolate natural functional RNAs guided by structural variation.

To compare generative properties, we used the regular DCA (REG) as our baseline, which we trained one instance per cluster, resulting in 25 $\times$ 2 REG models. We sampled from these models using $10^6$ MCMC steps that were initialized with the consensus sequence of the cluster, from which we next down sampled to 500 designs randomly. These designs typically reproduce well the distribution of the training data as shown in the PCAs in Fig. 4. However, these designs hardly generalize to the second cluster. This result is expected because the model did not have access to the second cluster sequences during the training.

To measure the coverage of the designed sequences, we used the PCA projection constructed with the natural sequences of each family. We first fitted an empirical distribution on each cluster in the PCA projection using the kernel density estimation (KDE) method (implemented in Sci-kit learn). We focused here on the principal components that represent 90\that defined a distribution for each cluster, we use Monte Carlo sampling to estimate the p-value of a design to be from one empirical distribution. For example, considering a 500 designs from a cluster A, we computed the p-value of each of the designs under the empirical KDE distribution of the second cluster B. Then, we defined as successful cross sampled RNAs the ones that have a p-value $> 0.01$, which means that they are not significantly different from the sequences in cluster B. Our results showed that REG allowed at least one successful cross sampled sequence in 14 pairs of clusters, for a total of 353 successful cross designs, as shown in Fig 4d. This result suggests that the clusters of these families are connected, meaning that a small number of mutations separate the two. Consistently, the average

base pair distance between consensus structures in these families is 3.

One intuitive strategy to enlarge the sampling coverage with MCMC is to increase the sampling temperature $T$. However, this approach did not improve the cross sampling success rate, see Fig 4d. On the contrary, the total number of successful cross sampling sequences dropped to 160. Similarly, introducing mutations with the biasing term $\omega$ to push the sampling away from the consensus sequence reduced the success of cross sampling to 217. The rational is that biasing the sampling from the training data without any guidance cannot improve the success rate, it only introduces uninformative noise.

Next, we test our hybrid (HYB) approach. To generate sequences with our hybrid model, we first trained it on one of the two sub MSA, for example the one corresponding to cluster A, using its cognate consensus structure. Then, we extracted the data-driven component, and we sampled sequences by combining it to the folding contribution computed using the consensus structure of the second cluster B. As above, we trained one instance of the model per cluster and extracted the data-driven part to cross sample sequences using the second consensus structure. To emphasis the role of the structure, we generated sequences with $\phi = 2$ and $\phi = 4$, from which we randomly extracted 500 designs. Ultimately, the goal is to leverage the opposite structure to steer the sampling toward the opposite cluster. Our results in Fig 4d show that HYB generated 624 sequences across all families, which corresponds to 15 pairs of clusters out of 50. We observed that the successful cross sampling obtained with HYB overlaps the ones obtained with the non structurally informed approach. However, the hybrid model significantly improved the yield of successful cross sampled RNAs, according to the Wilcoxon test with p-value=0.03. We show on Fig 4a,b two of the most successful cross designs we obtained with HYB.

While HYB doubled the success rate, it is not without limitations. The first caveat raised is that HYB still requires that the unseen cluster of sequences is partially overlapping the training point as observed above. The second caveat is related to the structure where it requires that the two consensus structures are slightly different ($\sim$ 3.4 base pairs distance), see Fig 4c. Identical pairs of consensus structures do not allow cross-sampling because it would only reproduce the training data, as REG. Finally, when the consensus structures are largely different ($> 10$ base pairs), our approach also failed at cross sampling.

Generating new but functional sequences beyond the training samples is a long-standing problem in ML. We evaluated our approach at recovering a hidden diversity. Our approach significantly improved the generation of sequences, suggesting that using the small structural perturbation to steer the generation beyond is an effective approach to explore new functional RNAs.

## Discussion

The RNA world hypothesis suggests that RNA once catalyzed reactions now performed by proteins, indicating that RNA has catalytic potential beyond what is currently observed in natural or experimental diversity. Exploring this potential requires designing functional RNAs that go beyond the diversity present in available data. Machine learning (ML) approaches are highly effective for designing RNA sequences; however, their reliance on training data constrains their ability to identify novel sequences. In contrast, biophysics-based methods are not bound by training data but have shown limited success in designing. This study bridges these two approaches, leveraging their respective strengths to design RNAs that extend beyond the constraints of training data.

The integration of a biophysical folding model into ML significantly improves RNA structure prediction, particularly for 3D contacts that are challenging to infer with traditional methods. These coevolutionary-based approaches, such as DCA, perform well in predicting 2D contacts because it is explicitly used to compute MSAs. However, they often fail to capture weaker, non-canonical interactions that are critical for RNA function. By explicitly incorporating folding stability into the training process, our hybrid model shifts focus from secondary structures to subtle non-trivial contacts, enhancing its ability to predict functionally relevant interactions. While the improvement in prediction is modest, it uniquely captures function-related interactions that no other model can infer, achieving this without additional data and solely through structural insights.

The level of biophysical realism implemented with the folding contribution is a limitation. Biological processes involving structured RNAs are typically sensitive to a repertoire of folds that a single RNA can adopt with a specific kinetic. Our results suggest that the hybrid model mainly removes the strong biases introduced

by the alignment procedure which accounts for a guessed secondary structure [20]. Therefore, introducing folding kinetics instead of static folding should improve the generalization, for which computationally efficient tools exist [**?**].

Designing new RNA molecules using ML generative models is limited because these models are typically trained by maximum likelihood to reproduce the observed training data. The intuition behind the hybrid model is to use the data agnostic biophysical model to guide the design in regions of the sequence space where training data is scarce. An earlier naive implementation of a hybrid RNA design approach showed promising results in diversifying a group I intron [15]; however, this model did not disentangle the two contributions. We explicitly introduced the structural contribution in the model during the learning to effectively disentangle the functional contribution from the molecular stability contribution, as we illustrated here with simulated binding data. The data-driven component absorbed the contribution from the binding because the folding stability is already explicitly modeled by the biophysical component. These independent sub models offers an unprecedented tool that allows the exploration of unseen sequence and structure regions we called structural imprinting. Structural imprinting consist in applying the function modeled by the data-driven component on an alternative structural configuration. We illustrated on natural RNA families that structural imprinting allows us to sample in a controlled manner unseen functional diversity. Our hypothesis is that the alternative conformation used to steer the sampling is adding constraints, *e.g.* new base pairing to be satisfied, which reduces the possible neighboring sequences to explore. Indeed, structural imprinting fails at using large structural differences because it requires large sequence modifications, which therefore pushes the sampling too far from the data.

Learning and generating diversity with limited training data is crucial for studying the origin of life, a singularity in the evolutionary timeline where direct observations are essentially unavailable. The hypothesized diversity of functional RNAs in the RNA world suggests that even the natural RNAs observed today can be diversified far beyond mere sequence variations. Our approach aims to recover this lost RNA diversity by leveraging structure to guide the generation of RNA molecules. However, it is currently limited to functions observed in natural RNAs. The next step would be to enhance the biophysical *a priori*, for example, by explicitly modeling desired catalytic functions with less coarse-grained structural representations (*e.g.* molecular mechanics [21]). Unfortunately, such approaches remain computationally intensive, making them incompatible with explicit RNA sampling, where structural contributions must be continuously evaluated. We anticipate that surrogate biophysical models, powered by deep neural networks, can overcome these limitations. Ultimately, hybrid models could enable the generation of functional RNAs with a significant success rate, including functions that are yet to be observed.

## References


[1] Rnacentral: a hub of information for non-coding rna sequences. *Nucleic Acids Research 47*, D1 (2019), D221–D229.

[2] Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC bioinformatics 9* (2008), 1–13.

[3] Calvanese, F., Lambert, C. N., Nghe, P., Zamponi, F., and Weigt, M. Towards parsimonious generative modeling of rna families. *Nucleic Acids Research 52*, 10 (2024), 5465–5477.

[4] Cao, L., Goreshnik, I., Coventry, B., Case, J. B., Miller, L., Kozodoy, L., Chen, R. E., Carter, L., Walls, A. C., Park, Y.-J., Strauch, E.-M., Stewart, L., Diamond, M. S., Veesler, D., and Baker, D. De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science 370*, 6515 (2020), 426–431.

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

[6] Crick, F. The origin of the genetic code. *Journal of Molecular Biology 38*, 3 (1968), 367–379.

[7] Cuturello, F., Tiana, G., and Bussi, G. Assessing the accuracy of direct-coupling analysis for rna contact prediction. *RNA 26*, 5 (2020), 637–647.

[8] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science 378*, 6615 (2022), 49–56.

[9] Dingle, K., Ghaddar, F., Šulc, P., and Louis, A. A. Phenotype bias determines how natural rna structures occupy the morphospace of all possible shapes. *Molecular biology and evolution 39*, 1 (2022), msab280.

[10] Fedor, M. J., and Williamson, J. R. The catalytic diversity of rnas. *Nature reviews Molecular cell biology 6*, 5 (2005), 399–412.

[11] Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology 35*, 2 (2017), 128–135.

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

[13] Jung, J. K., Alam, K. K., Verosloff, M. S., Capdevila, D. A., Desmau, M., Clauer, P. R., Lee, J. W., Nguyen, P. Q., Pastén, P. A., Matiasek, S. J., et al. Cell-free biosensors for rapid detection of water contaminants. *Nature biotechnology 38*, 12 (2020), 1451–1459.

[14] 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. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Research 49*, D1 (2021), D192–D200.

[15] Lambert, C. N., Opuu, V., Calvanese, F., Zamponi, F., Hayden, E., Weigt, M., Smerlak, M., and Nghe, P. Expanding the space of self-reproducing ribozymes using probabilistic generative models. *bioRxiv* (2024), 2024–07.

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

[17] Lu, X.-J., Bussemaker, H. J., and Olson, W. K. Dssr: an integrated software tool for dissecting the spatial structure of rna. *Nucleic acids research 43*, 21 (2015), e142–e142.

[18] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences 101*, 19 (2004), 7287–7292.

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

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

[21] Opuu, V., Nigro, G., Gaillard, T., Schmitt, E., Mechulam, Y., and Simonson, T. Adaptive landscape flattening allows the design of both enzyme: substrate binding and catalytic power. *PLOS Computational Biology 16*, 1 (2020), e1007600.

[22] Pucci, F., Zerihun, M. B., Peter, E. K., and Schug, A. Evaluating dca-based method performances for rna contact prediction by a well-curated data set. *RNA 26*, 7 (2020), 794–802.

[23] Sato, K., Akiyama, M., and Sakakibara, Y. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature communications 12*, 1 (2021), 941.

[24] SCHUG, A., WEIGT, M., ONUCHIC, J. N., HWA, T., AND SZURMANT, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences 106*, 52 (2009), 22124–22129.

[25] SUMI, S., HAMADA, M., AND SAITO, H. Deep generative design of rna family sequences. *Nature Methods* (2024), 1–9.

[26] SUN, H., ZHU, X., LU, P. Y., ROSATO, R. R., TAN, W., AND ZU, Y. Oligonucleotide aptamers: new tools for targeted cancer therapy. *Molecular Therapy-Nucleic Acids 3* (2014).

[27] **V. OPUU**, NIGRO, G., LAZENNEC-SCHURDEVIN, C., MECHULAM, Y., SCHMITT, E., AND SIMONSON, T. Redesigning methionyl-trna synthetase for $\beta$-methionine activity with adaptive landscape flattening and experiments. *Protein Science 32*, 9 (2023), e4738.

[28] **V. OPUU**, SUN, Y. J., HOU, T., PANEL, N., FUENTES, E. J., AND SIMONSON, T. A physics-based energy function allows the computational redesign of a pdz domain. *Scientific Reports 10*, 1 (2020), 11150.

[29] WATSON, J. L., JUERGENS, D., BENNETT, N. R., TRIPPE, B. L., YIM, J., EISENACH, H. E., AHERN, W., BORST, A. J., RAGOTTE, R. J., MILLES, L. F., ET AL. De novo design of protein structure and function with rfdiffusion. *Nature 620*, 7976 (2023), 1089–1100.

[30] YEN, L., SVENDSEN, J., LEE, J.-S., GRAY, J. T., MAGNIER, M., BABA, T., D'AMATO, R. J., AND MULLIGAN, R. C. Exogenous control of mammalian gene expression through modulation of rna self-cleavage. *Nature 431*, 7007 (2004), 471–476.

[31] ZERIHUN, M. B., PUCCI, F., AND SCHUG, A. Coconet—boosting rna contact prediction by convolutional neural networks. *Nucleic acids research 49*, 22 (2021), 12661–12672.

[32] ZHANG, S., LI, J., AND CHEN, S.-J. Machine learning in rna structure prediction: Advances and challenges. *Biophysical Journal* (2024).

[33] ZUKER, M., AND STIEGLER, P. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research 9*, 1 (1981), 133–148.

## Code and data availability

The code and data used in this research is available at https://github.com/vaiteaopuu/rna_design_biophy_ml. The training of the hybrid model finishes in seconds to minutes depending on the length of the MSA.

## Acknowledgments

I thank Philippe Nghe and Matteo Smerlak for helpful discussions. I thank Frederik Koch for doing the preliminary investigations.

## Figures

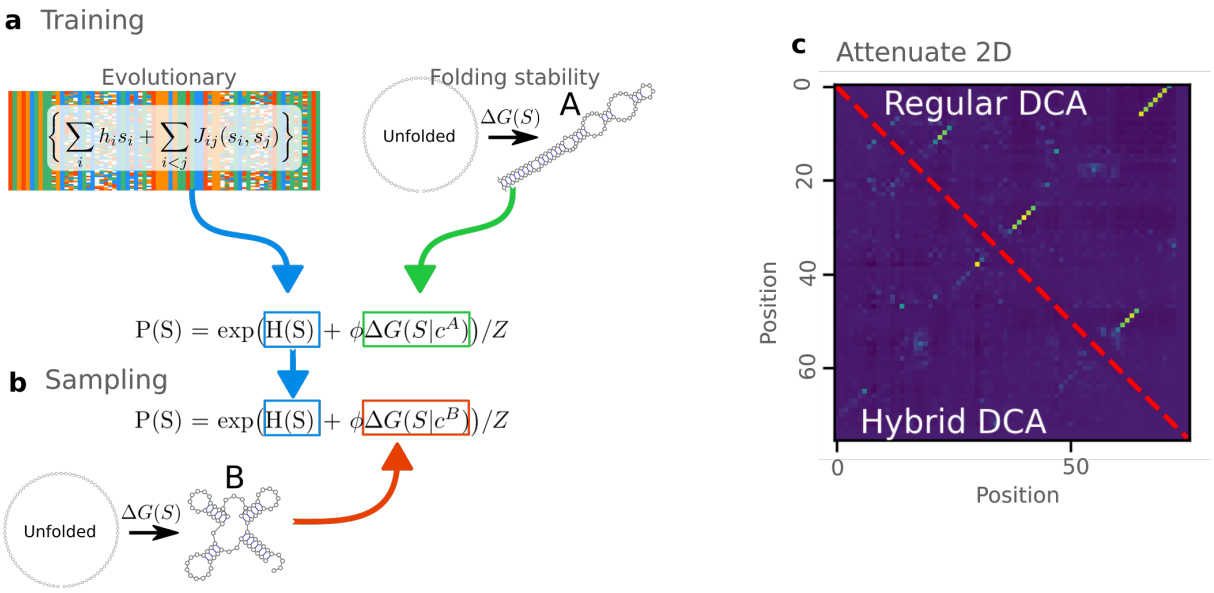


Figure 1: Hybrid DCA-folding framework. a) Training the hybrid model. The model composed of the DCA part and the folding part is trained on an MSA. The folding part requires a structural configuration that can be the MFE for example. The $\phi$ parameter is a trainable parameter that indicate the importance of the structure in the MSA. b) Sampling sequences with structural imprinting. To sample new sequences, we vary the structural configuration. c) Shifting the focus from 2D. We show here the coevolutionary scores (where higher is brighter) computed with the regular DCA (upper triangle) and the DCA part of the hybrid model (lower triangle) trained on the tRNA MSA.

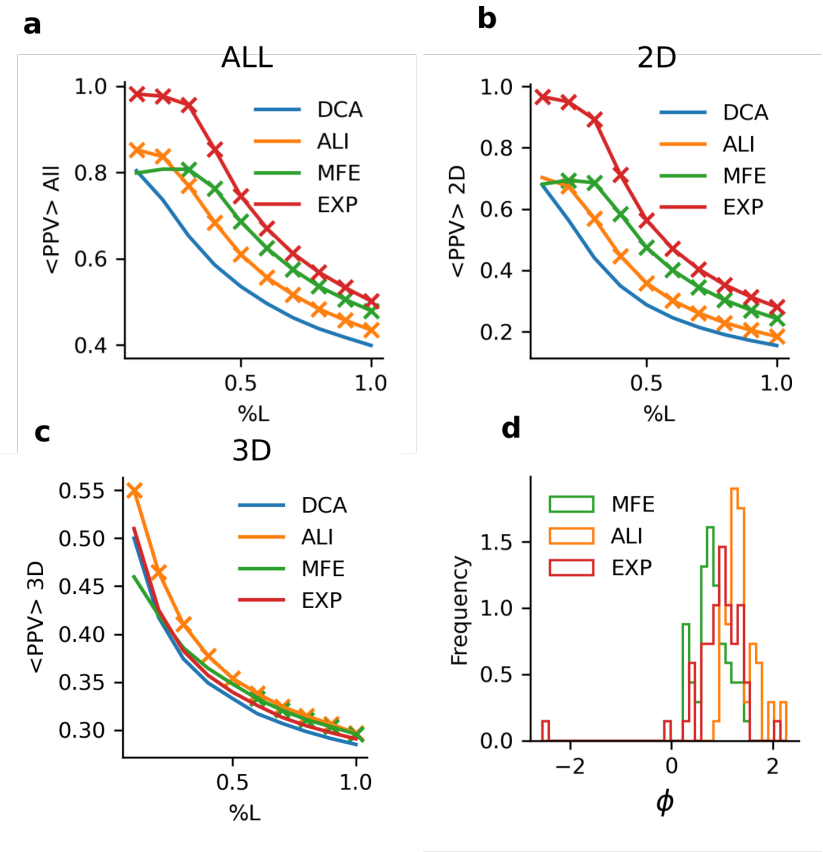


Figure 2: Prediction of non-trivial interactions. a-c) Prediction accuracy for ALL, 2D, and 3D contacts as a function of ranked contacts based on the scores $S_{i,j}$ for the regular DCA (blue), the hybrid DCA with the ALI configuration (ALI), the hybrid DCA with the MFE configuration (MFE), the hybrid DCA with the EXP configuration (EXP) models. The number of ranked contacts is expressed as a fraction of the RNA length (L) on the x-axis, while accuracy (y-axis) is measured as PPV. Crosses indicate the fraction of L values where the corresponding method significantly outperformed the regular DCA (in blue), as determined by the Wilcoxon test (pvalue $< 0.05$ ). d) Distribution of $\phi$ with hybrid models using MFE, ALI, or EXP configurations across all tested families.

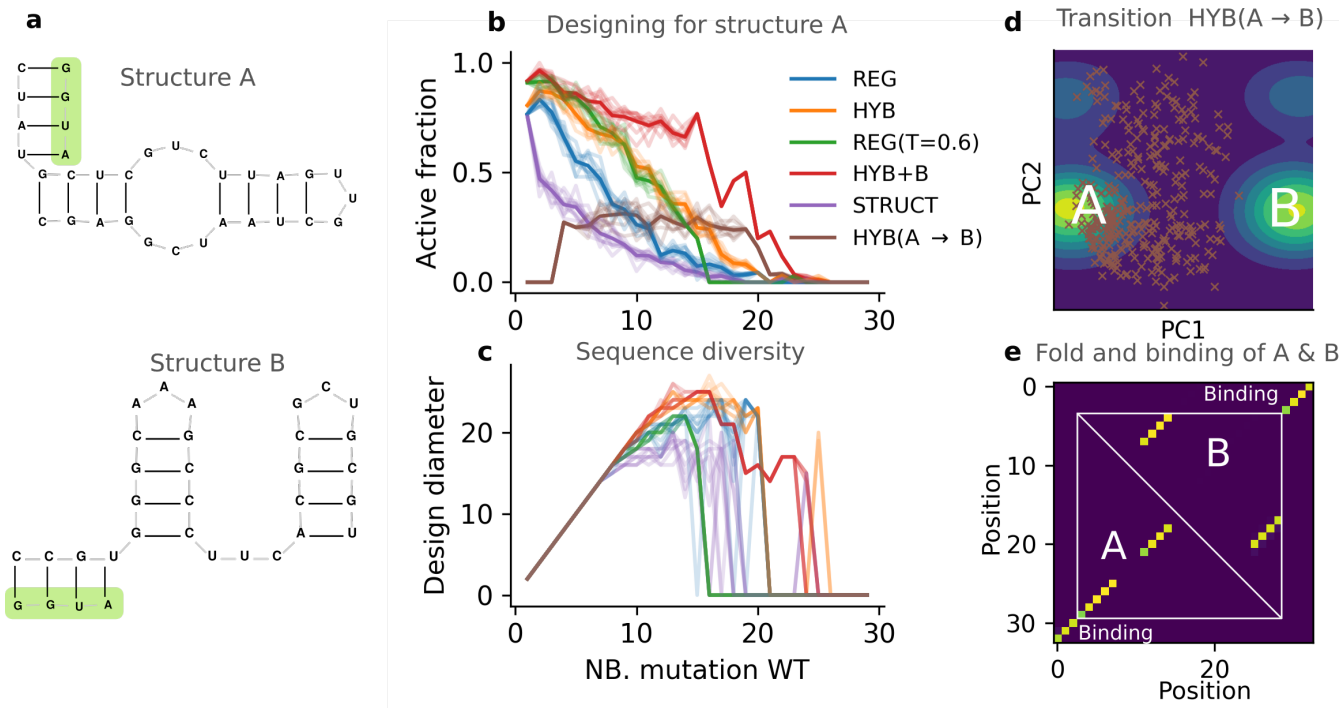


Figure 3: Disentanglement of stability from function. a) We show the secondary structure of the two configurations we considered where the ligand is represented in green. b) Success rate of design methods. The graph shows the fraction of active sequences designed as a function of their number of mutations to the wild type (WT). For each distance number of mutation to the WT (in x-axis), we randomly selected a 100 sequences that where designed by each method (REG=regular DCA; HYB=our hybrid model using configuration A; REG(T=0.6)=regular DCA at sampling lower sampling temperature; HYB+B=hybrid model with a stronger weight on the data-driven part during sampling; HYB(A $\rightarrow$ B)=our hybrid model using the configuration B). We repeated this process 10 times to display the variability of the experiment using shadowed lines. c) Sequence diversity. For each number of mutation to the WT, we recorded the number of mutation between the two most different designs, which we called design diameter. d) Design interpolation between binders in configuration A to binders in configuration B. We computed the PCA representation of sequences we artificially designed with random mutations of the respective WTs, which we represented as an annotated colored contour where bright regions means high density of sequences. We projected on the PCA of training sequences the designs we obtained with HYB(A $\rightarrow$ B) with brown crosses. e) Base pair probability of successfully designed sequences for both configuration.

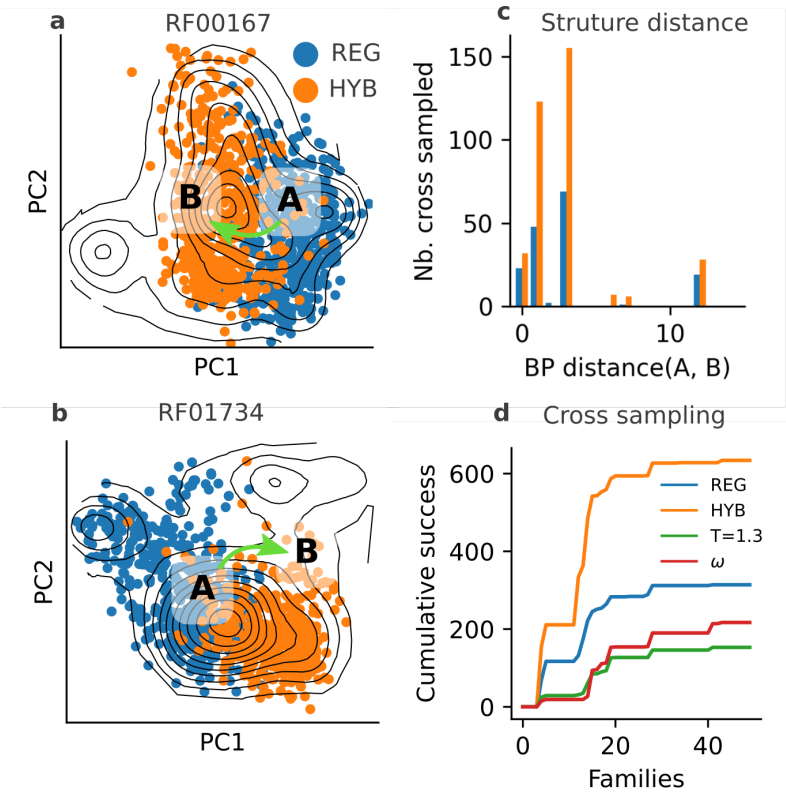


Figure 4: Generalization to unseen sequences. a-b) Structure imprinting to generalize to unseen cluster of sequences in families of natural RNAs. The scatter plots represents the PCA analysis of the natural RNAs with density contours representing the density of natural sequences. We annotated the center of the A and B clusters. The dots represent the projected designs with the REG and HYB methods where REG reproduces the training data (either A or B) and HYB uses the structural configuration of the second cluster to extrapolate the design. The green arrows represents the extrapolation direction. c) The number of successfully cross designed RNAs as a function of the base pair distance (the number of base pair that differ between the two configurations A and B) is represented in bar plots for the REG and HYB methods. d) Success of designing RNAs from the unseen cluster. For each family, we used the kernel density estimation to count how many designs fall in the cluster of unseen sequences under the pvalue threshold 0.01. We show the cumulative number of successfully cross sampled sequences across all 25 RNA families considered here.

# Methods

## Biophysically informed mean field DCA

Mean field DCA is a variant that uses a mean field approximation of the probability of sequences to estimate the coupling parameters [19]. It starts from the covariance matrix that is $C_{ij}(a,b) = f_{ij}(a,b) - f_i(a) * f_j(b)$ where at positions $i$ and $j$ with nucleotides $a$ and $b$, $f_{ij}(a,b)$ is the frequency of both nucleotides whereas $f_i(a)$ and $f_j(b)$ are the single frequencies. Then, the couplings are extracted from the inverse of the covariance matrix: $J_{ij}(a,b) = -(C^{-1})_{ij}(a,b)$.

Typically, the contribution of sequences to the frequencies are reweighted to limit the redundancy of the MSA [19]. Here, we follow the same principle and assume that the distribution is biased by the folding stability $\Delta G$. However, to account for the variation of energy distribution across RNA families, we centered and reduced the folding stability free energies per family, which we denote here $\Delta G(s|c)^*$. Thus, to compute the folding bias weight, we used the exponential normalization:

$$\frac{1}{w_s} = \frac{\exp(-\Delta G(s|c)^*)}{\sum_k \exp(-\Delta G(s|c)^*)}.$$

Then, the frequencies are computed as follows $f_i(a) = \sum_k s_k \delta_{a,s_k^i} \frac{1}{w_k}$, where the sum runs over all sequences of the MSA, $\delta$ is the Kronecker symbol that is 1 if the nucleotide $s_k^i$ of sequence $k$ at position $i$ is the same as $a$. We implemented this variant using Python Numpy package (version 1.21.6).

## Bias toward a reference sequence

Introducing mutations randomly in a functional natural RNA typically results in a an exponential loss of activity because most mutations are deleterious. To model explicitly the exponential decay of activity when introducing mutations from the wild type sequence, we introduce the bias term $\omega$:

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\frac{1}{T}(H(S) + \phi \Delta G(S|c))\right) + \omega D(S, \text{WT}),$$

where $D(S, WT)$ is the number of mutations between $S$ and the WT sequence. We update $\omega$ following the gradient:

$$\frac{\partial L}{\partial \omega} = \langle D(S, WT) \rangle_{MSA} - \langle D(S, WT) \rangle_{P(S,\omega)}.$$

Similarly to the update of $\phi$, we approximate the second term with the instantaneous $D(S, WT)$ of the sequence $S$ sample when the parameter is updated.

## Simulated folding binding mutational experiment

We describe here the generation of an artificial dataset for stability and binding from a single functional RNA. To generate the simulated dataset of an RNA, we considered a single-stem configuration we called A (see Fig 3), which binds to a ligand that is a small four-nucleotide RNA (with the sequence GGUA). To construct the initial functional RNA sequence, we performed MCMC sequence optimizations to find sequences that fold in the targeted structural configuration and binds the ligand sequence with the given pose. To select sequences, we used the structure-based cost function described in [15], which uses the McCaskill algorithm to compute the base pairing probability within the RNA receptor and with the ligand. This cost function ensures that the RNA does not fold in alternative configurations, and that the ligand does not bind in alternative poses. The resulting wild-type (WT) RNA (Fig. 3) was randomly mutated: for 1 to 10 mutations, we performed 10000 random mutation of the WT and selected the sequences with a structure based score $<$ 0.6. We obtained 342 mutants with at most 10 mutations, see Fig SI 7.

The binding is measured as the probability that each position of the binding site is correctly paired with the corresponding nucleotide in the ligand. This probability is also calculated using the McCaskill algorithm. Therefore, the binding is not the interaction free energy but depends on all the possible configurations between both partners, which means that the ligand binding score is greater if it specifically binds to the defined interaction site in 5'. A mutant is active for binding if the binding score is greater than 3.

## Figure SI

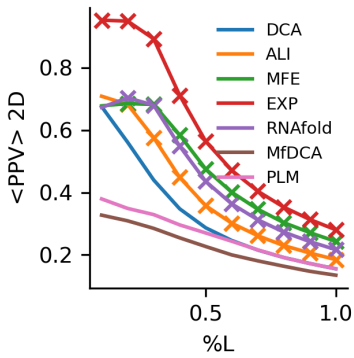


Figure 5: Prediction of non-trivial interactions. Prediction accuracy for 2D contacts as a function of ranked contacts based on the scores $S_{i,j}$. We record here the PPV for the regular DCA (blue), the hybrid DCA with the ALI configuration (ALI), the hybrid DCA with the MFE configuration (MFE), the hybrid DCA with the EXP configuration (EXP), the MFE structure predicted with RNAfold (RNAfold), the Mean Field DCA (MfDCA), and the pseudo-likelihood DCA (PLM). The number of ranked contacts is expressed as a fraction of the RNA length (L) on the x-axis, while accuracy (y-axis) is measured as PPV. Crosses indicate for that fraction of $L$, the PPV of the corresponding method significantly outperformed the regular DCA (in blue), as determined by the Wilcoxon test (pvalue $< 0.05$ ).

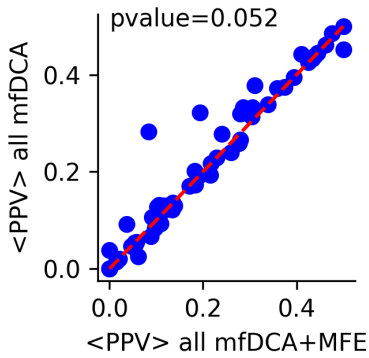


Figure 6: Comparison between biophysicall informed MfDCA and not informed in PPV.

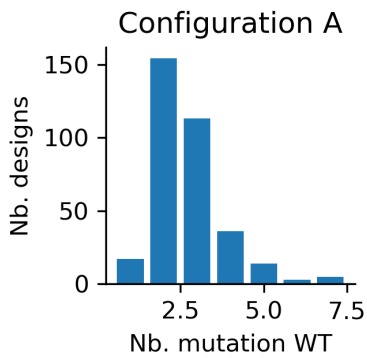


Figure 7: Artificial mutagenesis data for binding and stability. The barplot shows the number of we selected for folding and binding that are random mutants of the artificial WT sequence we obtained for the configuration A.