RNAtranslator: Modeling protein-conditional RNA design as sequence-to-sequence natural language translation

Sobhan Shukueian Tabrizi[1] Sina Barazandeh[2] Helyasadat Hashemi Aghdam[1]
A. Ercüment Çiçek[1*]
[1]Department of Computer Engineering, Bilkent University, Ankara, Türkiye 06800
[2]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Protein-RNA interactions are essential in gene regulation, splicing, RNA stability, and translation, making RNA a promising therapeutic agent for targeting proteins, including those considered undruggable. However, designing RNA sequences that selectively bind to proteins remains a significant challenge due to the vast sequence space and limitations of current experimental and computational methods. Traditional approaches rely on in vitro selection techniques or computational models that require post-generation optimization, restricting their applicability to well-characterized proteins.

We introduce RNAtranslator, a generative language model that formulates protein-conditional RNA design as a sequence-to-sequence natural language translation problem for the first time. By learning a joint representation of RNA and protein interactions from large-scale datasets, RNAtranslator directly generates binding RNA sequences for any given protein target without the need for additional optimization. Our results demonstrate that RNAtranslator produces RNA sequences with natural-like properties, high novelty, and enhanced binding affinity compared to existing methods. This approach enables efficient RNA design for a wide range of proteins, paving the way for new RNA-based therapeutics and synthetic biology applications. The model and the code is released at github.com/ciceklab/RNAtranslator.

1 Introduction

Protein-RNA binding is an important component of the interactome and plays crucial roles in gene expression regulation [1], splicing [2] , RNA stability [3], molecular degradation [4], RNA transport [5], and translation [6]. RNAs bind to proteins through either (i) RNA-binding domains (RBDs), which have rigid and well-defined 3D structures and selectively bind to specific RNA motifs, or (ii) Intrinsically Disordered Regions (IDRs), which have highly flexible structures, allowing them to interact dynamically with multiple binding RNA partners. The latter provides exciting opportunities for the design of novel RNA molecules that target and bind with proteins [7]. These can be used to inhibit, enhance, or modify RNA-protein interactions to restore functionality or eliminate disease-causing factors. RNA is a new and promising therapeutic agent to target proteins because only 15% of human proteins have binding pockets for ligands (small molecules) which are used as the common drug mechanism today [8]. This means that RNA has the potential to target

*Correspondence: cicek@cs.bilkent.edu.tr

proteins that have previously been deemed undruggable either with direct binding or by inhibiting the precursor RNAs. RNA-based drugs are also very favorable because (i) they have a lower immune system activation and inflammatory response with nucleoside-modifications; (ii) they have no genotoxicity risk as they do not integrate into the genome; (iii) they are relatively swift and inexpensive to produce with optimized chemistry and delivery [9].

The bottleneck in protein-conditional RNA design is the sheer number of possibilities for the candidate sequence, and unfortunately, current practice in the laboratory relies on manual and experimental design. The first strategy is to generate a large pool of candidate binding RNA sequences for a target using tools such as Primer-BLAST [10] followed by in vitro filtering of the library using experimental validation approaches such as SELEX (Systematic Evolution of Ligands by EXponential Enrichment) [11] which becomes more stringent in every validation cycle. The key element for success when it comes to selecting a good library to start with is based on luck, which is clearly not desirable. The second approach begins with an existing solution to a related problem, if one can be found, which is then optimized through the same in vitro test cycle. Both methods cost significant time and money and produce a suboptimal design. Computational techniques can aid design but due to their key limitations they are limited to proof-of-concept.

Most computational RNA design methods focus on *unconditional* RNA design. While early efforts solve the inverse folding problem via optimization [12, 13, 14, 15, 16, 17, 18, 19, 20], recent works use deep/reinforcement learning to design RNA structure (2D or 3D) [21, 22, 23, 24, 25, 26, 27] or RNA sequence [28, 29, 30, 31, 32]. Although the these methods can design synthetic RNA with natural-like properties such as GC content and minimum-free energy, only a few are capable of designing RNAs to target specific proteins (i.e., protein-conditional design). RNAFlow [24] is the only model that can design a *structure* to target a protein which uses a conditional flow-matching method. However, state-of-the-art structure designing models including RNAFlow cannot design novel molecules with variable length and the models can only recapitulate structures in the dataset [26]. This is because RNA is structurally very flexible compared to proteins and the paucity of ground truth structures bar training deep learning models [33]. Thus, the performances are random when designing conformationally diverse RNAs [24].

Protein-conditional RNA *sequence* design models do not suffer from training data scarcity and learn a latent distribution over widely available RNA sequence databases using architectures like generative adversarial networks [28, 30] and GPT-based language models [31]. However, conditioning on the target requires a post-generation optimization step with a binding affinity prediction method for the target protein. This means the current state-of-the-art procedure have key bottlenecks: (1) The two-step approach (generation+optimization) does not harness the full potential of learning better embeddings for RNA in relation to other molecules (targets); (2) The pipeline is dependent on the availability of RNA interaction data for the target and/or third-party tools that predict the binding affinity of the designed RNA with the target protein. For example, such DeepBind [14] models are available for only 732 proteins and DeepCLIP models [34] are available for 221 proteins. This means we cannot design RNA to target under-studied proteins or novel synthetic proteins. This is a crucial bottleneck for the applicability of this system as the number of proteins in the human body is estimated to be between 20,000 to several millions depending on the definition [35, 36, 37].

In this study, we propose a novel approach to overcome the above-mentioned challenges and enable designing binding RNA for any target protein for the first time. That is, for the first time, we formulate the problem of protein-conditional RNA design as a sequence-to-sequence *natural lan-*

guage translation problem (not to be confused with biological process of protein generation). We use large-scale RNA-protein interaction datasets to train a generative language model, RNAtranslator, end-to-end. Like ChatGPT inputing a sentence in English and translating it into French, RNAtranslator inputs a target protein sequence and *translates* it to a novel binding RNA sequence. The model learns a latent space for both the RNA and the protein along with their interactions and thus, does not require any postprocessing or optimization method to tailor the design for binding to the target. Thus, the model is capable of designing RNA to bind with any target protein sequence.

Our results show that RNAtranslator is able generate RNAs which resemble natural binding RNAs with respect to length, GC content, minimum free energy (MFE) and ensemble free energy distributions, yet we also find that the designed sequences are novel and diverse. We observe that RNAtranslator is substantially better compared to state-of-the-art methods and is capable of designing for any target protein. Despite being stable, we also find that the designed RNAs have higher binding affinity to the target compared to the state-of-the-art using binding affinity and docking prediction methods. We think the potential of RNAtranslator goes beyond this study and it will pave the way for modeling and designing for many other interactions and will be the first milestone in the new field of *molecular translation*. The model and the code is released at github.com/ciceklab/RNAtranslator.

2 Results

2.1 RNAtranslator Overview

RNAtranslator is an encoder-decoder transformer-based language model designed for RNA sequence generation conditioned on a target protein sequence input. The model aims to generate RNA sequences that exhibit strong binding affinity to target protein while maintaining structural stability and natural RNA-like properties. The architecture of RNAtranslator is shown in Figure 1. The model is trained end-to-end with 26 million pairs of protein-RNA interactions obtained from RNA interaction databases. The model encoder is input with the target protein sequence. The decoder inputs the binding RNA sequence and the embedded protein sequence and learns to regenerate the RNA sequence from left to right. During inference, the model inputs the target protein sequence and generates novel RNA sequences that are likely to bind to it. We select the candidate sequence after sampling from this step and filtering on the basis of metrics such as binding affinity and stability. To evaluate our designs, we use in silico binding affinity predictors such as DeepCLIP [34] and docking tools to simulate the molecular interaction between the design and the target structure.

We compare our approach against natural RNAs and RNAs designed by state-of-the-art methods using RBM5 protein as our target which is an RNA-binding protein. We obtain the list of RNA molecules that are listed as binding to this protein in the RNA-protein interaction databases. We call list of 128 RNAs the *Natural Binding RNAs*. We randomly select the same number of natural RNAs and call this list *Natural Non-Binding RNAs*. Then, we design 128 RNAs to target RBM5 using RNAtranslator and compare the distributions of these sets with respect to several stability metrics mentioned above. We also generate the same number of RNAs to target RBM5 using two state-of-the-art methods and compare them with RNAtranslator. These models are RNAGEN [28] and GenerRNA [31].

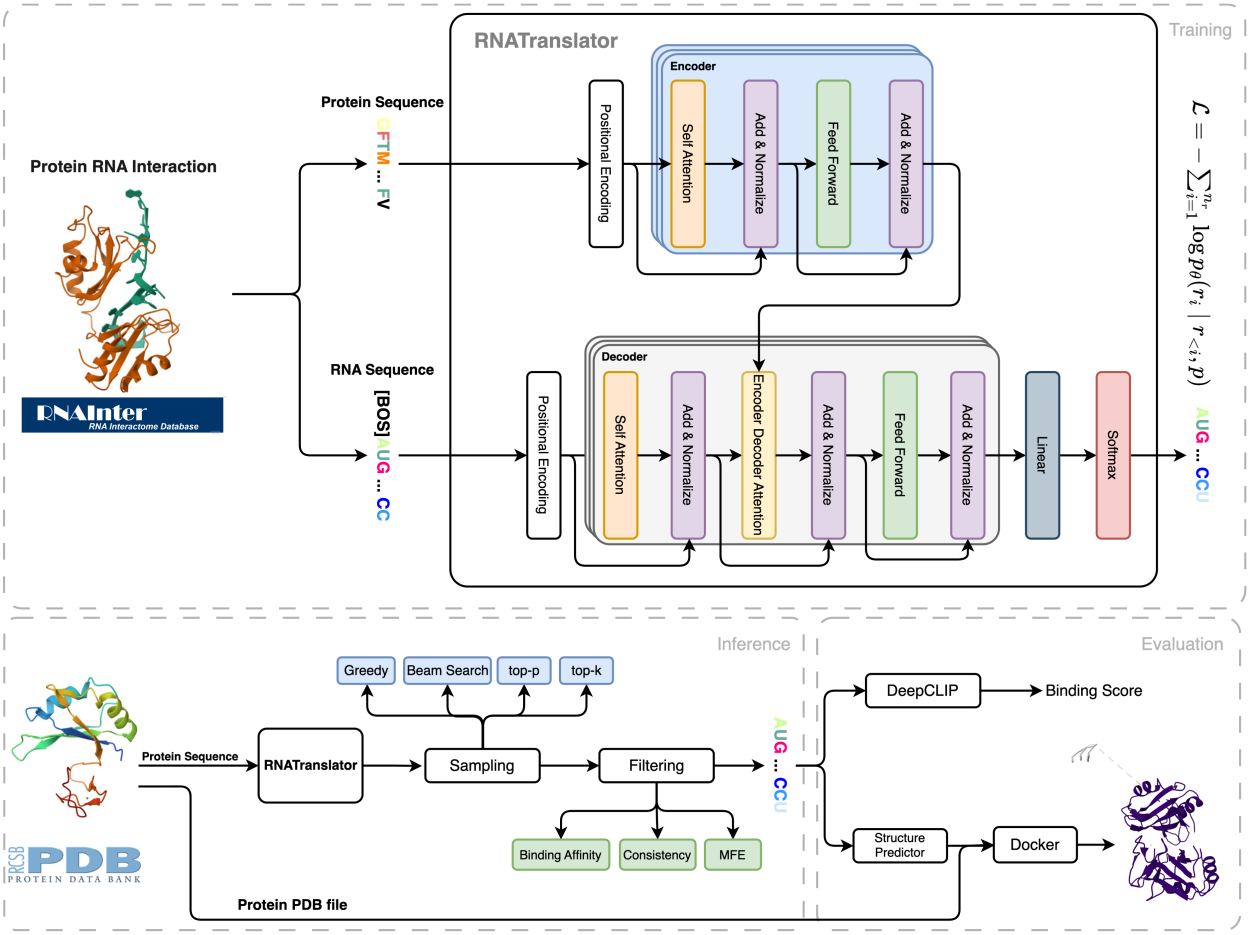


Figure 1: **Overview of RNAtranslator pipeline:** The model follows an encoder-decoder transformer framework, where the encoder processes the protein sequence using positional encoding and self-attention layers to extract an embedding. The decoder takes the RNA sequence input, applying self-attention, encoder-decoder attention, and feed-forward layers to learn to generate target-specific RNA sequences. During inference only the protein sequence is input and model designs a novel RNA sequence after a sequence of sampling and filtering procedures to ensure high binding affinity and stability. We evaluate the design based on in-silico predicted binding affinity to the target and molecular simulations of the complex using docking tools.

2.2 RNAtranslator generates novel RNAs

To evaluate the similarity between RNAtranslator-generated RNAs and natural RBM5 binding RNA sequences, we employed BLAST (Basic Local Alignment Search Tool), a widely used algorithm for sequence alignment. This method allowed us to systematically compare our 128 generated RNAs with known binding RNAs used in training. To determine whether a sequence is novel or not, we applied specific thresholds: Sequences with percent identity ≥ 85, alignment length $\geq 10bp$, and E-value $\leq 1e-3$ are classified as not novel, while those falling below these thresholds are considered novel. Our analysis revealed that only 9 percent of the generated RNAs exhibit high similarity with known sequences, indicating that RNAtranslator predominantly generates novel RNA sequences rather than replicating existing ones.

2.3 RNAtranslator-designed RNAs are stable

To evaluate the stability of RNAtranslator-generated RNA sequences, we analyzed key stability metrics, including free energy, GC content, and ensemble free energy unfolding. The Minimum Free Energy (MFE) serves as an indicator of RNA stability by predicting the most thermodynamically favorable secondary structure. Lower MFE values suggest more stable conformations, as they represent the minimum energy required for an RNA molecule to fold [38]. Furthermore, GC content was examined as a complementary measure, since a higher GC content enhances structural robustness through stronger base-pairing and higher melting temperatures [39].

Our results show that RNAtranslator-generated RNAs are stable with respect to these metrics. The analysis of GC content in Figure 2B shows that natural binding RNAs tend to have a larger GC content compared to nonbinding natural RNAs, as expected. RNAs designed by RNAtranslator and GenerRNA have even a stronger GC composition compared to natural binding RNAs. In contrast, RNAGEN-generated sequences tend to have lower GC content similar to non-binding RNAs, suggesting a potential reduction in stability. Our analysis on MFE also further supports these findings. As illustrated in Figure 2C, RNAtranslator generated RNAs achieve a distribution slightly more stable than GenerRNA and more stable than natural or RNAGEN-generated RNAs. This suggests that RNAs generated with RNAtranslator have a better energy landscape, which enhances their overall foldability and thermodynamic resilience. To provide a more comprehensive assessment of RNA stability, we incorporate the free energy unfolding of the ensemble ($\Delta G_{ensemble}$), which considers the entire thermodynamic ensemble of possible secondary conformations rather than just the lowest energy structure using ViennaRNA[40]. As shown in Figure 2D, RNAtranslator generated RNAs have a distribution that tend to be slightly more negative than GenerRNA-generated RNAs. The length of the RNA sequence also affects the secondary structure and binding affinity. It is also an important factor for diversity and novelty of the design. For this reason, we also compare the RNA length distributions across different groups. As shown in Figure 2A, RNAtranslator-generated RNAs span a similar range like GenerRNA and natural binding RNAs with higher median compared to natural non-binding RNAs while RNAGEN can generate shorter sequences with a lower variance.

Overall, compared to other methods, RNAtranslator and GenerRNA generated RNAs display a more stable energy distribution with a large span of sequence lengths where RNAtranslator has a slight advantage. This reinforces RNAtranslator's ability to adopt biologically relevant folding patterns under diverse conditions. We would like to point out to one advantage of our method that we needed to optimize the RNAs generated by RNAGEN using a binding affinity predictor trained

specifically for RBM5 and we had to fine-tune GenerRNA model with RNAs that are known to bind to RBM5. RNAtranslator is free from these steps which enables it to generalize to any protein sequence and can work in the absence of these information.

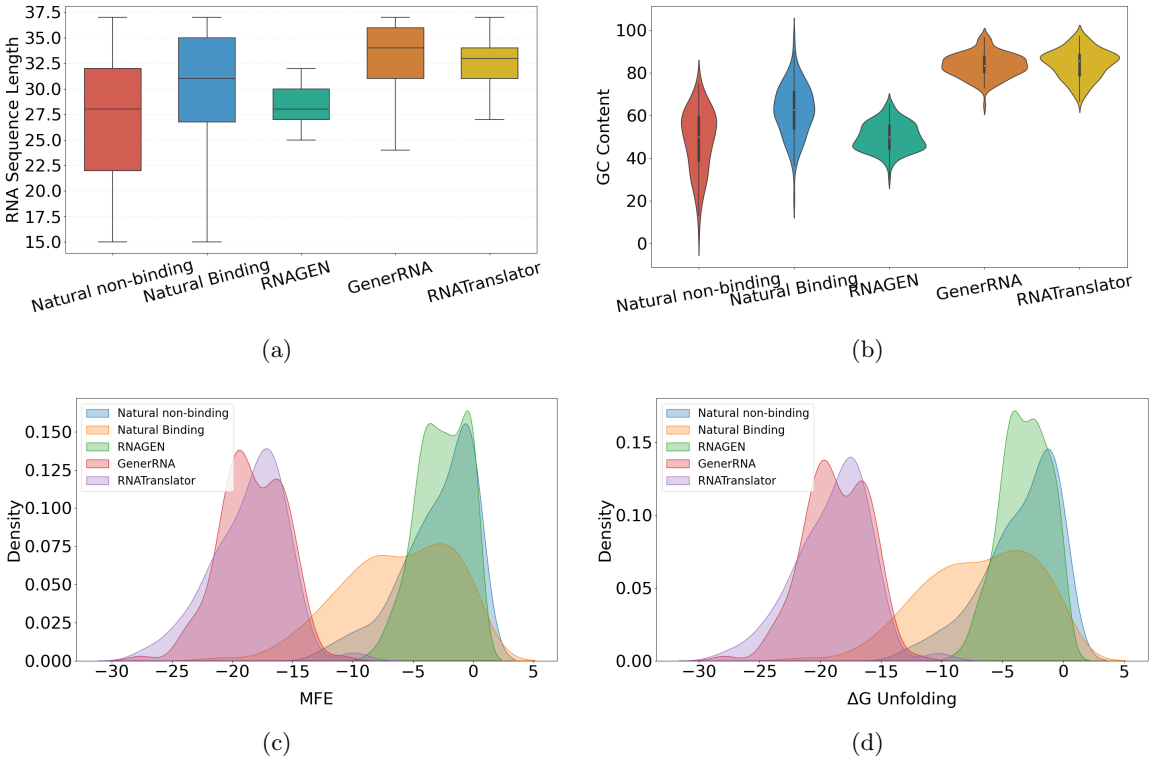


Figure 2: **Stability Analysis of RNAtranslator-Generated RNAs(a)** Distribution of RNA sequence lengths across different groups. **(b)** GC content comparison across different groups. **(c)** Minimum Free Energy (MFE) distributions across different groups. **(d)** Ensemble Free Energy ($\Delta G_{ensemble}$) distributions across different groups.

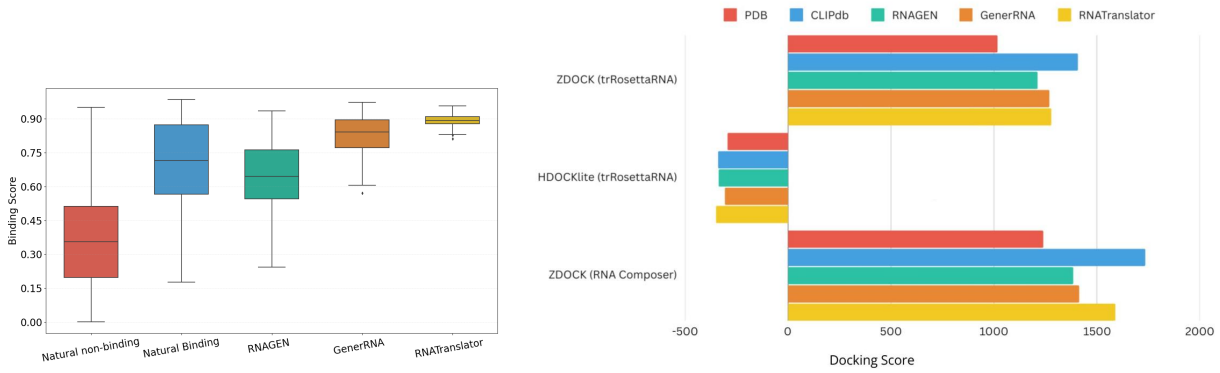
2.4 RNAtranslator can design RNAs with high binding affinity to the target

We evaluate the binding affinity of the designed RNAs using both deep learning-based interaction prediction models and docking simulations. We use the same sets of RNAs from Section 2.3. First, we use the DeepCLIP model[34] to quantify the RNA-protein binding affinity given a protein and RNA sequence. This model predicts the probability of binding based on features that correlate with known protein-binding motifs and RNA sequence patterns. In Figure 3A, we show the binding affinity distributions for each set of RNAs. RNAtranslator-generated RNAs have substantially higher median binding scores compared to other groups of RNAs. In addition, the binding affinities of RNAtranslator maintain a lower variance, suggesting that our designs have higher confidence compared to others.

We also analyze the designed RNAs at the structural level and simulate RNA-protein complexes using docking simulations. Since ground-truth structure for the designed RNAs are not available, we predict the three-dimensional structures using two methods to ensure robustness and minimize

potential structural inaccuracies. First, we use RNA Composer which is a template-free modeling tool that constructs 3D RNA structures based on sequence and known structural motifs [41]. Then, we use TrRosettaRNA, which is a deep learning-based RNA structure prediction method that predicts 3D RNA structure by leveraging evolutionary information and deep neural networks trained on RNA structural constraints [42]. We obtain the ground-truth 3D protein structure from the PDB database [43]. Using these structures, we conduct docking simulations with ZDOCK and HDOCKLite to evaluate the RNA-protein binding potential. ZDOCK is a fast Fourier transform-based rigid-body docking method that predicts molecular interactions by exploring possible binding conformations and scoring them based on shape complementarity, electrostatic potential, and desolvation energy [44]. Meanwhile, HDOCKlite is a lightweight version of the HDOCK docking algorithm that focuses on RNA-protein interactions, incorporating a scoring function optimized for molecular docking [45] By these docking approaches, we evaluated how well the generated RNA sequences interact with their target protein, providing a binding score and potential protein RNA complex structure.

We compare the ZDOCK and HDOCKlite binding scores of the best RNAtranslator-generated RNA (i.e., top predicted binding affinity to the target with respect to DeepCLIP) with (i) a natural binding RNA to RBM5 obtained from PDB [46], (ii) a natural binding RNA to RBM5 obtained from CLIPdb [47] and (iii) designed top RNAs from state-of-the-art methods. We use 3D predicted structures of the RNAs using two methods mentioned above. As shown in Figure 3B, RNAtranslator-generated RNAs achieve substantially higher docking scores than the state-of-the-art methods using both trRosettaRNA and RNAComposer as structure prediction methods, highlighting their better structural compatibility with target proteins. Notably, in HDOCKLite, lower docking scores indicate stronger binding affinity, and designed RNA achieve the best score, even better than experimentally-validated natural binding RNAs.



(a) DeepCLIP predicted binding affinities to RBM5

(b) Docking Scores of the top designs and natural binding RNAs

Figure 3: **Binding Affinity Analysis of RNAtranslator-Generated RNAs. (A)** Binding scores predicted using DeepCLIP for different RNA groups. RNAtranslator-generated RNAs exhibit higher and more consistent binding scores compared to other methods and natural RNAs. **(B)** ZDock and HDOCKLite docking scores for protein-RNA complexes using different structure prediction methods (trRosettaRNA and RNA Composer). RNAtranslator-generated RNAs achieve high docking scores, indicating strong structural compatibility with target proteins, comparable to CLIPdb and PDB-derived natural binding RNAs.

Table 1: Structural validation metrics for different RNA-RBM5 complexes obtained using natural binding RNAs obtained from PDB and CLIPdb and the top synthetic RNAs generated by each method.

	PDB	**CLIPdb**	**RNAGEN**	**GenerRNA**	**RNAtranslator**
Probably wrong sugar puckers	20.00%	2.50%	0.00%	0.00%	0.00%
Bad backbone conformations	50.00%	12.50%	3.85%	50.00%	0.00%

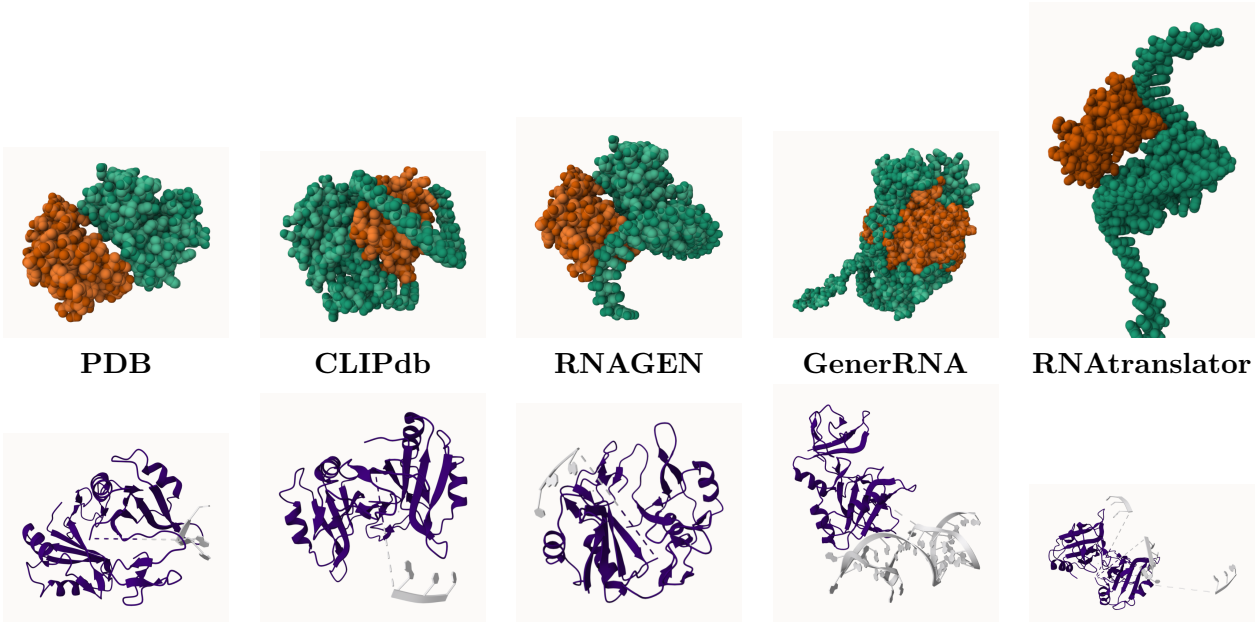


Figure 4: **Structural Comparison of Natural and Computationally Generated RNA-Protein Complexes:** The figure compares RNA-protein complexes using ground-truth natural binding RNAs (from PDB and CLIPdb) and RNAs generated by different computational methods (RNAGEN, GenerRNA, and RNAtranslator). The top row shows space-filling models, with green representing the RNA molecule and orange representing proteins. The bottom row uses cartoon representations to emphasize backbone conformations, showing protein folding in purple and RNA in a lighter gray.

The 3D visualization of the RNA-protein complexes obtained using ZDOCK are shown in Figure 4. We further analyze the quality of the RNA-protein complexes by validating their structural integrity using MolProbity, a widely used structure validation tool for biomolecules [48]. MolProbity assesses atomic-level correctness by identifying steric clashes, unfavorable backbone conformations, and geometric distortions in macromolecular structures. This evaluation ensures that our predicted complexes are computationally favorable and structurally sound. Specifically, we compare the generated complexes using the following metrics: (i) *Probably wrong sugar puckers*, which measures the percentage of incorrect pucker conformations adopted by the sugar backbone, deviating from the expected geometry; and (ii) *Bad backbone confirmations*, which measures the percentage of energetically unstable spatial arrangements of the phosphate-sugar backbone. As shown in Table 1, the RNAtranslator-generated RNA-protein complexes exhibit no geometric irregularities with

respect to these metrics and obtain better results than even natural binding RNAs.

2.5 Molecular dynamics simulations show RNAtranslator can design RNAs that form stable complexes with the target

We use OpenMM [49] to simulate the dynamics of our design. Molecular dynamics simulations help reveal how these complexes behave in a physiologically relevant environment rather than a static crystal structure. These simulations provide insights into conformational changes, hydrogen-bonding networks, solvent effects, and the thermodynamics of binding at atomic resolution by using force fields optimized for both proteins and RNA. Please see Section 4.6 for the details on the simulation. We analyze the Free Energy Landscape (FEL) of the design which describes the energy landscape of a molecule or a complex as a function of atoiv coordinates and provides insights into the stability, folding, binding interactions, and conformational dynamics.

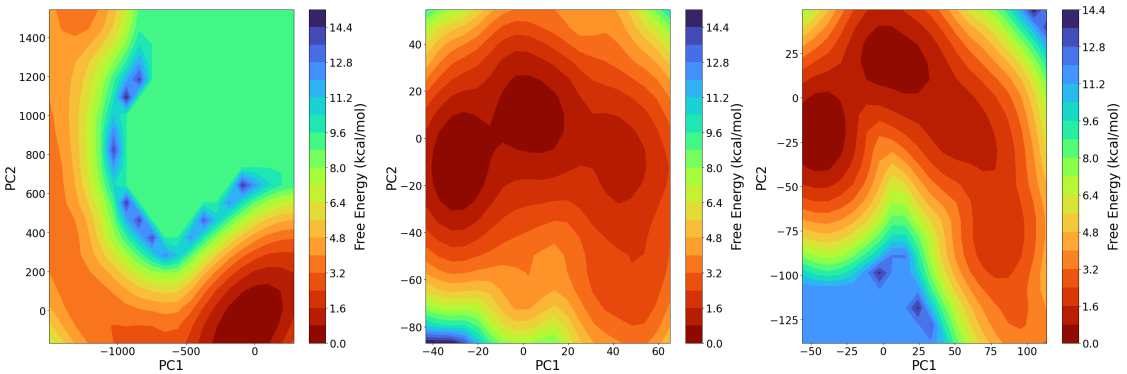
We show the the first two principal components of FELs in Figure 5 for (i) our top designed RNA to target RBM5; (ii) a natural binder RNA of RBM5 and (iii) a randomly generated RNA with 37nucleotides. These are shown in the corresponding rows. Columns show (i) FEL for the RNA when it is isolated, and (ii) FEL for only RNA but when RNA and RBM5 are bound, and (iii) the FEL for the RNA-RBM5 complex.

We observe that all three RNAs span a broader conformational distribution when isolated as the coordinate ranges for the PCs in column 1 are an order of magnitude wider than those for columns 2 and 3. We observe that the random RNA has a very different conformational distribution compared to our top design and the ground-truth binder whereas the latter are almost identical.

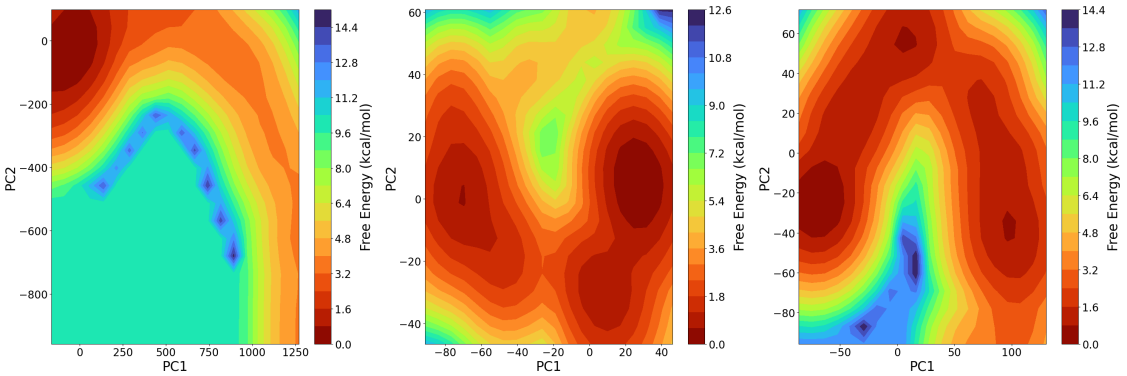
Upon binding to RBM5, the FEL for the natural binder RNA and the designed RNA become more confined, reflecting a reduction in conformational heterogeneity and suggesting the formation of stable, well-defined complexes as reflected in narrower and similar PC coordinate ranges (Rows 1 & 2 and Columns 2 & 3). Although high-energy areas (blue) are very seldom in these four subplots, for the randomly generated RNA this region covers a larger percentage of the plot (Row 3 and Columns 2 & 3). Again, the FEL of the complex with the designed RNA closely resembles the FEL of the natural binder RNA and protein complex. These results suggest that our top design resembles a natural RNA and dynamically shows similar patterns as a natural binder to this protein while being very diverse sequence-wise.

3 Discussion

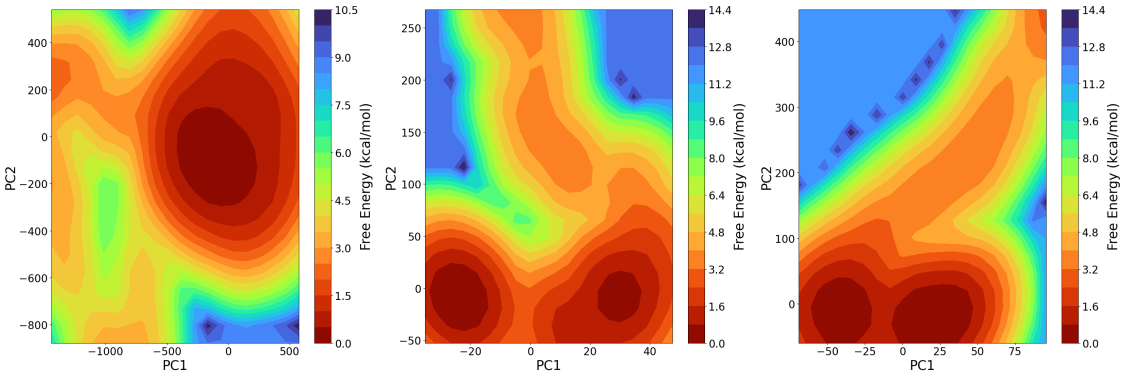
We introduce a novel approach to protein-conditional RNA sequence design by formulating the problem as a natural language translation task. Unlike existing methods that rely on separate generation and optimization steps, our model, RNAtranslator, learns a unified latent space for RNA and protein interactions, enabling direct and target-specific RNA design. The results demonstrate that RNAtranslator outperforms state-of-the-art approaches in the generation of natural-like RNA sequences while improving binding affinity to the target. Our results suggest that the designed RNAs maintain stability while preserving sequence and length diversity. This balance is important, as excessive stability may lead to undesired rigidity which bars effective binding, while high flexiblity may lack specificity in the design. Additionally, our docking and binding affinity evaluations confirm that RNAtranslator designs natural-like RNA sequences with superior binding potential compared to existing models.



(a) FELs for the RNAtranslator's top designed RNA to bind RBM5



(b) FELs for a natural binder RNA for RBM5



(c) FELs for a randomly generated RNA for RBM5

Figure 5: **Free energy landscape (FEL) Analysis of RNAtranslator-Generated RNAs Compared to a Natural Binder and Random RNAs.** Each figure has three columns: The first shows the FEL of the RNA when isolated (unbound), the second shows the FEL of the RNA when bound to RBM5, and the third shows the FEL of the RNA - RBM5 complex (bound). **(A)** RNA designed by our model for RBM5, and **(B)** naturally binding RNA to RBM5, and **(C)** a randomly generated RNA sequence. Unbound RNA show high diversity, but upon binding, both the natural and designed RNAs form narrow, low-energy basins indicative of stable complexes, with the designed RNA closely mirroring or exceeding the stability of the natural RNA.

The advantage and the novelty of RNAtranslator lies in its ability to generalize beyond predefined RNA-protein interactions. Traditional RNA design methods rely on either (1) experimental selection processes, which are costly and time-intensive, or (2) computational models that suffer from data scarcity and the need for post-generation optimization. We eliminate these bottlenecks by using large-scale RNA-protein interaction datasets and end-to-end training on paired data, allowing the model to learn implicit binding rules without manual intervention. Moreover, RNAtranslator does not rely on third-party binding affinity predictors, which are rarely available only for a few hundred well-studied proteins. For instance, we had to optimize the RNAs generated by RNAGEN using a binding affinity predictor trained specifically for RBM5 and we had to fine-tue the GenerRNA method with RNAs that are known to bind to RBM5. RNAtranslator does not require any of these steps as it was traine with large scale interaction data, end-to-end. This allows RNAtranslator to expand the scope of RNA-based therapeutics.

Despite these promising results, there are several areas for further improvement. One challenge is the scalability of training large language models on biological sequences, given the computational complexity of RNA-protein interactions. Future work could explore techniques such as parameter-efficient fine-tuning or self-supervised learning on larger, unlabeled datasets to further enhance performance. Another limitation is the lack of experimentally validated binding data for novel RNA designs. While our computational assessments suggest high-affinity interactions, experimental validation through techniques such as surface plasmon resonance or isothermal titration caliometry will be necessary to confirm binding efficacy in vitro.

In summary, our study presents a paradigm shift in protein-conditional RNA design by introducing an end-to-end generative approach that eliminates the need for post-optimization or fine-tuning. RNAtranslator successfully designs RNA sequences with enhanced stability, diversity, and binding affinity, overcoming key challenges in existing computational and experimental methods. Our findings suggest that this framework will play a pivotal role in the future of RNA therapeutics, molecular engineering, and synthetic biology.

4 Methods

4.1 Dataset

We obtain the RNA-Protein interaction data from the RNAInter Database [50]. We use a total of 26 million experimentally-validated and computationally-predicted RNA-protein interactions (RPIs) from various sources, with a significant portion of its dataset focused on human interactions. The database includes interactions derived from high-throughput experimental techniques such as CLIP-seq, RIP-seq, and yeast three-hybrid assays, providing detailed annotations for different RNA types, including mRNA, lncRNA, miRNA, and circRNA.

4.2 Problem Formulation

Let $p = (p_1, \ldots, p_{n_p})$ denote a protein sequence, where each p_i is a one or multiple amino acids token, and let $r = (r_1, \ldots, r_{n_r})$ be the corresponding RNA sequence we aim to generate. Our primary objective is to learn the conditional distribution $p(r)$, which describes how likely an RNA sequence r is given a protein sequence p. By formulating this as a language translation problem, where the protein sequence is considered the source language and the RNA sequence is the target language,

we used techniques from natural language processing (NLP) to model conditional protein-RNA design. We factorize $p(r)$ via the chain rule:

$$p(r) \;=\; \prod_{i=1}^{n_r} p(r_i \mid r_{<i}, p) \tag{1}$$

We then train a model with parameters θ that minimizes the negative log-likelihood (NLL) over a protein-RNA interaction dataset $\mathcal{D} = \{(p^{(1)}, r^{(1)}), \ldots, (p^{(|\mathcal{D}|)}, r^{(|\mathcal{D}|)})\}$.

$$\mathcal{NLL}(\mathcal{D}) \;=\; -\sum_{k=1}^{|\mathcal{D}|} \sum_{i=1}^{n_r^{(k)}} \log\, p_\theta\big(r_i^{(k)} \mid r_{<i}^{(k)},\, p^{(k)}\big). \tag{2}$$

4.3 Model Architecture

We implement the RNAtranslator using the T5 architecture [51], a widely adopted encoder-decoder transformer designed for sequence-to-sequence tasks.

4.3.1 Training

We feed the encoder with the protein sequence p to model the protein language and learn the information from a protein's RNA-binding domains (RBDs). Before processing, each amino acid token is converted into a learned token embedding, using BPE tokenization to handle large protein vocabularies. Positional information is also encoded, enabling the model to differentiate between tokens in different sequence positions. Concretely, a tokenized protein sequence of length n_p is mapped to a sequence of embeddings, which is then passed through l layers of self-attention and feedforward blocks. The final output of the encoder is a matrix $H \in R^{n_p \times d}$ where d is the model dimension and encodes contextual information about the protein's amino acid composition, potential RNA-binding domains, and other functionally relevant features. Given the encoder output H, the decoder generates the target RNA sequence $(r_1, \ldots, r_{n_r})$ one token at a time. At each timestep i, the decoder conditions on, (i) the partially generated RNA sequence $(r_1, \ldots, r_{i-1})$ via *self-attention*, (ii) the encoder output H via *cross-attention*:

$$D_i = \mathrm{Dec}_\theta\big((r_1, \ldots, r_{i-1}), H\big) \tag{3}$$

Here, D_i denotes the decoder's hidden representation at timestep i. This vector $D_i \in R^d$ captures contextual information from both the evolving RNA sequence and the encoded protein features. Finally, a linear layer $W_{\text{vocab}} \in R^{d \times |\mathcal{V}|}$ projects D_i onto the vocabulary space $\mathcal{V}$ of RNA tokens, and a softmax function normalizes these scores to yield a probability distribution over all possible next tokens.

4.3.2 Inference

At inference time, we produce a *de novo* RNA sequence $\hat{r}$ for a given protein p by sampling tokens sequentially from the output distribution:

$$\hat{r}_1 \sim p_\theta(r_1 \mid p), \quad \hat{r}_2 \sim p_\theta(r_2 \mid \hat{r}_1, p), \quad \ldots \quad \hat{r}_{n_r} \sim p_\theta(r_{n_r} \mid \hat{r}_{<n_r}, p). \tag{4}$$

There are various techniques for sampling in language models [52, 53, 54]. For our work, we employ multiple approaches. One approach is greedy sampling, which selects the token with the highest probability at each time step. While this method often produces coherent sequences with high overall likelihood, it can be overly conservative and may lead to repetitive outputs. To mitigate this, we also use beam search. Beam search [52] maintains multiple candidate sequences by exploring the top B tokens (where B denotes the beam size) at each decoding step, eventually selecting the sequence with the highest overall probability. Although beam search generally improves sequence quality, it may still limit diversity. To introduce greater diversity, we incorporate stochastic sampling techniques such as top-k sampling and nucleus sampling (also known as top-p sampling). In top-k sampling [53], the token selection is restricted to the k most probable tokens, thereby reducing the likelihood of sampling low-probability tokens that could reduce sequence quality. Nucleus sampling, on the other hand, dynamically determines the candidate set based on a cumulative probability threshold p, ensuring that only the tokens contributing to the majority of the probability mass are considered. Both stochastic methods provide a balance between randomness and determinism, enabling the generation of RNA sequences that are diverse yet biologically plausible.

Recognizing that no single sampling strategy perfectly balances diversity and accuracy, we generate a pool of candidate RNA sequences using various sampling strategies and hyperparameter settings. We then apply filtering using predicted binding affinity, consistency, and minimum free energy (MFE). All of these metrics are normalized between zero and one, to rank the generated sequences effectively. We normalize the MFE using min-max scaling $MFE_{norm} = 1 - \frac{MFE - \min(MFE_{gen})}{\max(MFE_{gen}) - \min(MFE_{gen})}$ by this we ensure that lower MFE values are ranked higher. Binding affinity is derived from DeepCLIP model[34] and consistency measures the frequency of a given RNA sequence appearing across different sampling runs, normalized by dividing by the maximum occurrence observed $C_{norm} = \frac{C}{\max(C_{gen})}$. Higher consistency values indicate more robust sequence generation across sampling strategies.

4.4 Experimental Setup

We trained RNAtranslator on two NVIDIA TITAN RTX GPUs using Fully Sharded Data Parallel (FSDP) to split the training across devices efficiently. we tokenize RNA and protein sequences using the Byte Pair Encoding (BPE) algorithm. A vocabulary size of 1,000 tokens was chosen to balance representation quality and computational efficiency. Each sequence is truncated or padded to a fixed length of 1,024 tokens to ensure uniformity during model training. We set a batch size of 8 and used gradient accumulation over 16 steps. We also applied weight decay with the Adam optimizer. The training carried on for 350,000 iterations over a month. Our hardware configuration used FP16 with optimization level O1 and a fixed seed of 42 for reproducibility. The model has a stack of 6 identical encoder-decoder layers with 12 attention heads. The total number of parameters is 41.4 million.

We used the same machine for GenerRNA. We obtained the base model and fine-tuned the model with 3917 RNA interactions from CLIPdb database for RBM5 protein. We fine-tuned for 50,000 iterations. For RNAGEN, we used the released pre-trained model and generated 128 sequences which are optimized to bind to RBM5 using the DeepBind model [14] for the corresponding protein.

4.5 Evaluation Metrics

Our primary objective is to ensure that the RNAs generated by RNAtranslator closely resemble natural RNAs in both structural properties and binding characteristics. To achieve this, we first define a set of properties that capture stability (e.g., minimum free energy, predicted secondary structures) and binding specificity (e.g., affinity scores). We then compare the *distributions* of these properties across four groups of RNA sequences: (i) **Natural non-binding RNAs** which randomly selected RNAs that are not known to bind the target protein, (ii) **Natural binding RNAs** that experimentally validated RNAs that bind to the target protein, (iii) **RNAtranslator-generated RNAs** are produced by our proposed Binding RNA Language Model, and (iv) **Other generated RNAs** are generated by alternative computational design methods for comparison. To ensure a fair and comprehensive evaluation, we validate the generated RNAs by comparing their distribution.

Specifically, we select 128 RNAs from CLIPdb[47] as natural binding RNAs and choose the best 128 RNAs from the generated RNA pool. This selection process ensures that not only the highest-ranked RNA sequences are considered but also that the overall distribution of generated RNAs aligns well with natural binding RNAs. Furthermore, beyond sequence-based validation, we perform structural analysis to confirm that RNAtranslator generated RNAs are not only computationally favored but also structurally suitable for stable RNA-protein interactions. By examining the 3D conformations of these RNAs, we assess their feasibility in forming biologically relevant structures that align with known natural binding RNAs.

4.6 Molecular Dynamics Simulations

We use molecular dynamics simulation framework OpenMM [49] to assess structural dynamics of the designed RNAs. During the simulations, we employ a canonical ensemble (NVT) setup with a Langevin integrator at 300 K and a 2 fs time step. Each system undergoes energy minimization, followed by equilibration and a 10-million-step production run. Trajectories are recorded at regular intervals to track conformational changes over time. This standardized protocol is applied to all systems, including RNA–protein complexes and isolated RNAs.

Before running simulations, we preprocess the structures with pdbfixer[55] to correct missing or nonstandard residues and remove unnecessary heterogens, then prepare using the AMBER14 force field, ensuring appropriate protonation states at pH 7.0.

To analyze the conformational landscapes of our simulated systems, we use a principal component analysis (PCA)-based method to generate free energy landscapes (FELs)[56]. We process trajectory data with MDAnalysis[57], focusing on RNA and protein atom coordinates. Each frame is aligned to a reference structure to eliminate global translational and rotational motions. PCA is applied to the concatenated Cartesian coordinates, reducing complex motion to the two most relevant collective coordinates (PC1 and PC2). We then construct a two-dimensional histogram of these projections and apply Gaussian smoothing to reduce noise. The free energy is computed using Eq5.

$$F(\mathrm{PC}_1, \mathrm{PC}_2) = -k_\mathrm{B} T \ln P(\mathrm{PC}_1, \mathrm{PC}_2) \tag{5}$$

where k_B is the Boltzmann constant, and the temperature is set to 300 K. Here, $P(\mathrm{PC}_1, \mathrm{PC}_2)$ represents the probability distribution of the system's conformations in the principal component space. It is obtained by constructing a histogram of the sampled molecular dynamics trajectory

projected onto PC1 and PC2. Essentially, P quantifies how frequently the system visits particular conformational states during the simulation. Since free energy F is related to the probability of a state via the Boltzmann factor, regions of high probability correspond to low free energy, while less frequently sampled regions correspond to higher free energy.

The energy surface is normalized so that the minimum energy state is 0 kcal/mol, allowing for a clear visualization of energy basins that correspond to stable conformational states. This method provides insights into the dynamic stability and transitions of molecules.

References

[1] M. D. Díaz-Muñoz and M. Turner, "Uncovering the role of rna-binding proteins in gene expression in the immune system," *Frontiers in immunology*, vol. 9, p. 1094, 2018.

[2] Y. Tao, Q. Zhang, H. Wang, X. Yang, and H. Mu, "Alternative splicing and related rna binding proteins in human health and disease," *Signal Transduction and Targeted Therapy*, vol. 9, no. 1, p. 26, 2024.

[3] S. Srikantan and M. Gorospe, "Hur function in disease," *Frontiers in bioscience (Landmark edition)*, vol. 17, p. 189, 2012.

[4] A. Hasan, C. Cotobal, C. D. Duncan, and J. Mata, "Systematic analysis of the role of rna-binding proteins in the regulation of rna stability," *PLoS genetics*, vol. 10, no. 11, p. e1004684, 2014.

[5] M. P. Thelen and M. J. Kye, "The role of rna binding proteins for local mrna translation: implications in neurological disorders," *Frontiers in molecular biosciences*, vol. 6, p. 161, 2020.

[6] R. F. Harvey, T. S. Smith, T. Mulroney, R. M. Queiroz, M. Pizzinga, V. Dezi, E. Villenueva, M. Ramakrishna, K. S. Lilley, and A. E. Willis, "Trans-acting translational regulatory rna binding proteins," *Wiley Interdisciplinary Reviews: RNA*, vol. 9, no. 3, p. e1465, 2018.

[7] A. Balcerak, A. Trebinska-Stryjewska, R. Konopinski, M. Wakula, and E. A. Grzybowska, "Rna–protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity," *Open biology*, vol. 9, no. 6, p. 190096, 2019.

[8] M. G. Costales, J. L. Childs-Disney, H. S. Haniff, and M. D. Disney, "How we think about targeting rna with small molecules," *Journal of medicinal chemistry*, vol. 63, no. 17, pp. 8880–8900, 2020.

[9] Y.-K. Kim, "Rna therapy: rich history, various applications and unlimited future prospects," *Experimental & Molecular Medicine*, vol. 54, no. 4, pp. 455–465, 2022.

[10] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden, "Primer-blast: a tool to design target-specific primers for polymerase chain reaction," *BMC bioinformatics*, vol. 13, pp. 1–11, 2012.

[11] C. Chai, Z. Xie, and E. Grotewold, "Selex (systematic evolution of ligands by exponential enrichment), as a powerful tool for deciphering the protein–dna interaction space," *Plant transcription factors: methods and protocols*, pp. 249–258, 2011.

[12] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, P. Schuster *et al.*, "Fast folding and comparison of rna secondary structures," *Monatshefte fur chemie*, vol. 125, pp. 167–167, 1994.

[13] J. A. Garcia-Martin, P. Clote, and I. Dotu, "Rnaifold: a constraint programming algorithm for rna inverse folding and molecular design," *Journal of bioinformatics and computational biology*, vol. 11, no. 02, p. 1350001, 2013.

[14] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nature biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[15] A. Churkin, M. D. Retwitzer, V. Reinharz, Y. Ponty, J. Waldispühl, and D. Barash, "Design of rnas: comparing programs for inverse rna folding," *Briefings in bioinformatics*, vol. 19, no. 2, pp. 350–358, 2018.

[16] A. Esmaili-Taheri and M. Ganjtabesh, "Erd: a fast and reliable tool for rna design including constraints," *BMC bioinformatics*, vol. 16, pp. 1–11, 2015.

[17] Á. Rubio-Largo, L. Vanneschi, M. Castelli, and M. A. Vega-Rodríguez, "Multiobjective metaheuristic to design rna sequences," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 1, pp. 156–169, 2018.

[18] G. Minuesa, C. Alsina, J. A. Garcia-Martin, J. C. Oliveros, and I. Dotu, "Moirnaifold: a novel tool for complex in silico rna design," *Nucleic acids research*, vol. 49, no. 9, pp. 4934–4943, 2021.

[19] N. S. Merleau and M. Smerlak, "arnaque: an evolutionary algorithm for inverse pseudoknotted rna folding inspired by lévy flights," *BMC bioinformatics*, vol. 23, no. 1, p. 335, 2022.

[20] H. K. Wayment-Steele, D. S. Kim, C. A. Choe, J. J. Nicol, R. Wellington-Oguri, A. M. Watkins, R. A. Parra Sperberg, P.-S. Huang, E. Participants, and R. Das, "Theoretical basis for stabilizing messenger rna through secondary structure design," *Nucleic acids research*, vol. 49, no. 18, pp. 10 604–10 617, 2021.

[21] J. K. Franke, F. Runge, R. Köksal, R. Backofen, and F. Hutter, "Rnaformer: A simple yet effective deep learning model for rna secondary structure prediction," *bioRxiv*, pp. 2024–02, 2024.

[22] F. Runge, D. Stoll, S. Falkner, and F. Hutter, "Learning to design rna," *arXiv preprint arXiv:1812.11951*, 2018.

[23] C. Tan, Y. Zhang, Z. Gao, H. Cao, S. Li, S. Ma, M. Blanchette, and S. Z. Li, "R3design: deep tertiary structure-based rna sequence design and beyond," *Briefings in Bioinformatics*, vol. 26, no. 1, p. bbae682, 2025.

[24] D. Nori and W. Jin, "Rnaflow: Rna structure & sequence design via inverse folding-based flow matching," *arXiv preprint arXiv:2405.18768*, 2024.

[25] C. K. Joshi, A. R. Jamasb, R. Viñas, C. Harris, S. Mathis, A. Morehead, R. Anand, and P. Liò, "grnade: Geometric deep learning for 3d rna inverse design," *bioRxiv*, 2024.

[26] R. Anand, C. K. Joshi, A. Morehead, A. R. Jamasb, C. Harris, S. V. Mathis, K. Didi, B. Hooi, and P. Lio, "Rna-frameflow for de novo 3d rna backbone design," in *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

[27] H. Huang, Z. Lin, D. He, L. Hong, and Y. Li, "Ribodiffusion: tertiary structure-based rna inverse folding with generative diffusion models," *Bioinformatics*, vol. 40, no. Supplement_1, pp. i347–i356, 2024.

[28] F. Ozden, S. Barazandeh, D. Akboga, S. S. Tabrizi, U. O. S. Seker, and A. E. Cicek, "Rnagen: A generative adversarial network-based model to generate synthetic rna sequences to target proteins," *bioRxiv*, pp. 2023–07, 2023.

[29] S. Barazandeh, F. Ozden, A. Hincer, U. O. S. Seker, and A. E. Cicek, "Utrgan: Learning to generate 5'utr sequences for optimized translation efficiency and gene expression," *bioRxiv*, pp. 2023–01, 2023.

[30] A. T. Riley, J. M. Robson, and A. A. Green, "Generative and predictive neural networks for the design of functional rna molecules," *bioRxiv*, 2023.

[31] Y. Zhao, K. Oono, H. Takizawa, and M. Kotera, "Generrna: A generative pre-trained language model for de novo rna design," *PLoS One*, vol. 19, no. 10, p. e0310814, 2024.

[32] Z. Wang, Z. Liu, W. Zhang, Y. Li, Y. Feng, S. Lv, H. Diao, Z. Luo, P. Yan, M. He *et al.*, "De novo design and optimization of aptamers with aptadiff," 2023.

[33] C. Tan, Y. Zhang, Z. Gao, B. Hu, S. Li, Z. Liu, and S. Z. Li, "Rdesign: Hierarchical data-efficient representation learning for tertiary structure-based rna design," in *The Twelfth International Conference on Learning Representations*, 2024.

[34] A. G. B. Grønning, T. K. Doktor, S. J. Larsen, U. S. S. Petersen, L. L. Holm, G. H. Bruun, M. B. Hansen, A.-M. Hartung, J. Baumbach, and B. S. Andresen, "Deepclip: predicting the effect of mutations on protein–rna binding with deep learning," *Nucleic acids research*, vol. 48, no. 13, pp. 7099–7118, 2020.

[35] P. Gaudet, P.-A. Michel, M. Zahn-Zabal, A. Britan, I. Cusin, M. Domagalski, P. D. Duek, A. Gateau, A. Gleizes, V. Hinard *et al.*, "The nextprot knowledgebase on human proteins: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D177–D182, 2017.

[36] R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia *et al.*, "How many human proteoforms are there?" *Nature chemical biology*, vol. 14, no. 3, pp. 206–214, 2018.

[37] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund *et al.*, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, 2015.

[38] M. Zuker, "On finding all suboptimal foldings of an rna molecule," *Science*, vol. 244, no. 4900, pp. 48–52, 1989. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.2468181

[39] N. Galtier and J. Lobry, "Relationships between genomic g+ c content, rna secondary structures, and optimal growth temperature in prokaryotes," *Journal of molecular evolution*, vol. 44, pp. 632–636, 1997.

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

[41] M. Biesiada, K. J. Purzycka, M. Szachniuk, J. Blazewicz, and R. W. Adamiak, "Automated rna 3d structure prediction with rnacomposer," *RNA Structure Determination: Methods and Protocols*, pp. 199–215, 2016.

[42] W. Wang, C. Feng, R. Han, Z. Wang, L. Ye, Z. Du, H. Wei, F. Zhang, Z. Peng, and J. Yang, "trrosettarna: automated prediction of rna 3d structure with transformer network," *Nature Communications*, vol. 14, no. 1, p. 7266, 2023.

[43] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[44] B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng, "Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers," *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, 2014.

[45] Y. Yan, H. Tao, J. He, and S.-Y. Huang, "The hdock server for integrated protein–protein docking," *Nature protocols*, vol. 15, no. 5, pp. 1829–1852, 2020.

[46] S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, and S. Velankar, "Protein data bank (pdb): the single global macromolecular structure archive," *Protein crystallography: methods and protocols*, pp. 627–641, 2017.

[47] Y.-C. T. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, and Z. J. Lu, "Clipdb: a clip-seq database for protein-rna interactions," *BMC genomics*, vol. 16, pp. 1–8, 2015.

[48] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "Molprobity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 1, pp. 12–21, 2010.

[49] P. Eastman, R. Galvelis, R. P. Peláez, C. R. Abreu, S. E. Farr, E. Gallicchio, A. Gorenko, M. M. Henry, F. Hu, J. Huang *et al.*, "Openmm 8: molecular dynamics simulation with machine learning potentials," *The Journal of Physical Chemistry B*, vol. 128, no. 1, pp. 109–116, 2023.

[50] J. Kang, Q. Tang, J. He, L. Li, N. Yang, S. Yu, M. Wang, Y. Zhang, J. Lin, T. Cui *et al.*, "Rnainter v4. 0: Rna interactome repository with redefined confidence scoring system and improved accessibility," *Nucleic acids research*, vol. 50, no. D1, pp. D326–D332, 2022.

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[52] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[53] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," *arXiv preprint arXiv:1805.04833*, 2018.

[54] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[55] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern *et al.*, "Openmm 7: Rapid development of high performance algorithms for molecular dynamics," *PLoS computational biology*, vol. 13, no. 7, p. e1005659, 2017.

[56] A. Sabei, C. Hognon, J. Martin, and E. Frezza, "Dynamics of protein–rna interfaces using all-atom molecular dynamics simulations," *The Journal of Physical Chemistry B*, vol. 128, no. 20, pp. 4865–4886, 2024.

[57] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domanski, D. L. Dotson, S. Buchoux, I. M. Kenney *et al.*, "Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations," Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), Tech. Rep., 2019.