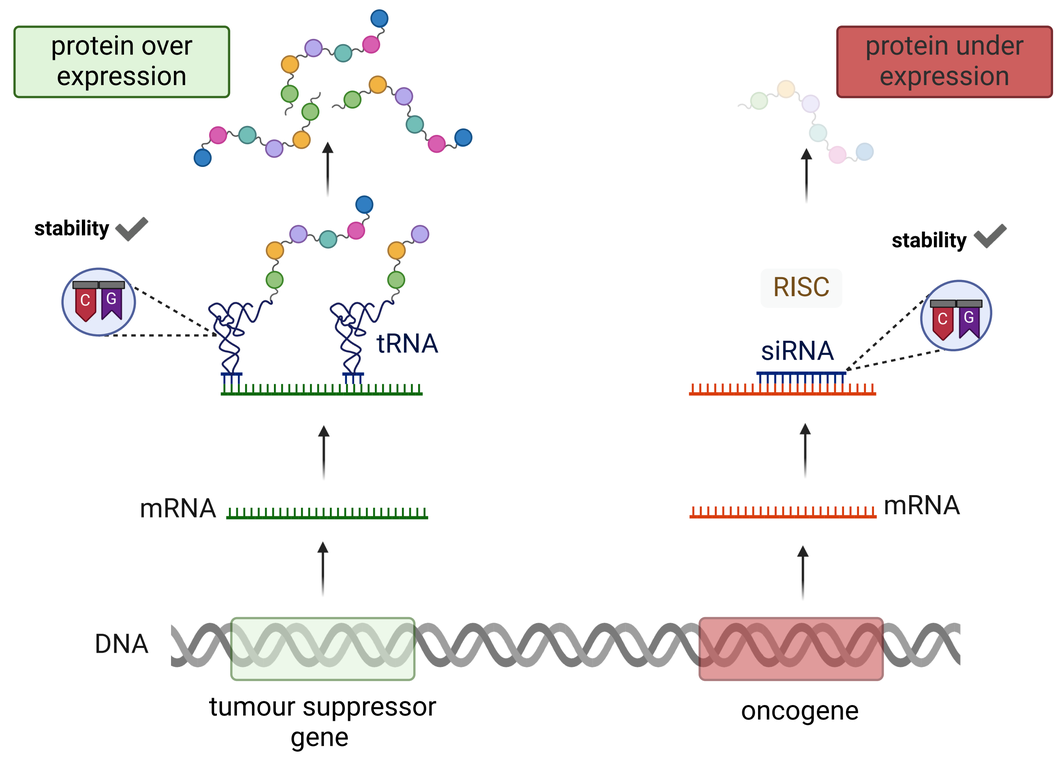
Deep diffusion model to generate ncRNAs

Mathieu Dujardin, Alexandre Dessagne

CentraleSupélec

RNA, a key molecule for life and why making it more stable through deep learning

- RNA (ribonucleic acid) is a crucial biomolecule involved in various biological processes, including gene expression, regulation, and protein synthesis. It is a single-stranded molecule made up of nucleotides, each consisting of a Ribose sugar, a Phosphate group a Nitrogenous base: Adenine (A), Uracil (U), Cytosine (C), and Guanine (G). While the most well-known type of RNA, messenger RNA, acts as a template for protein synthesis, a significant portion of them is non-coding. These RNA molecules do not code for proteins but are vital for regulating and maintaining cellular processes. For example, tRNA delivers amino acids to ribosomes enabling protein synthesis, while siRNA silence genes by preventing protein synthesis.
- The aim of this study is to use generative AI to generate new ncRNAs sequences, in the process of creating new and possible sequence we will aim to optimize the sequence stability by optimising the G/C content, the proportion of those 2 nucleotides greatly affect the molecule stability. In case of tumour suppressor gene having a more stable tRNA helps to create more healing proteins, in case of an oncogene, having a more stable siRNA can help creating genic therapy.



RNA-FM : biologically-meaningfull embedding

RNA FM is a BERT like(encoder-only) language model based on transformers. It is trained on a large database oof both coding and non-coding RNAs. This model captures a biologically-meaningful representation for each sequence. The model needs to take into account both each nucleotide (because C and G are what will be looked at to make sure the sequence is stable) but also take into account the context. For an RNA sequence: AUGGCUAC, each nucleotide is a encoded as a token. The representation of G in GGCU will reflect not just the identity of G but also its structural and functional relationship with the neighboring nucleotides G, C, and U.

Study of the Latent Space of RNA-FM

This is the model we used in practice to explore clustering patterns among RNA categories. Indeed, we analyzed human RNA sequences from RNAcentral, a comprehensive database of non-coding RNA sequences, using RNA-FM. The sequences were extracted from a FASTA file, categorized into broad RNA types (e.g., long non-coding RNA, microRNA, antisense RNA ...), and transformed into high-dimensional embeddings. We then applied t-SNE to visualize clustering patterns and explore RNA similarity.

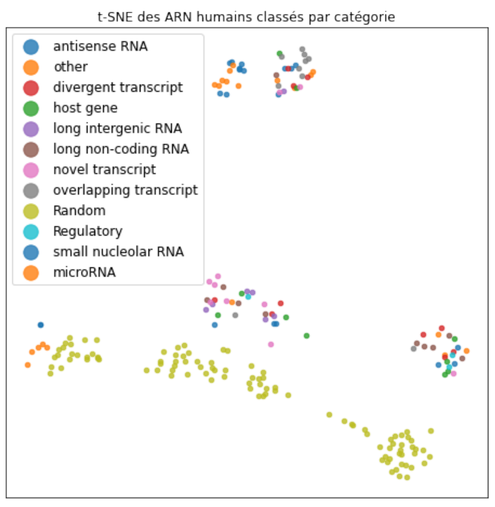


Figure 1. t-SNE of human RNA sequences categorized by type

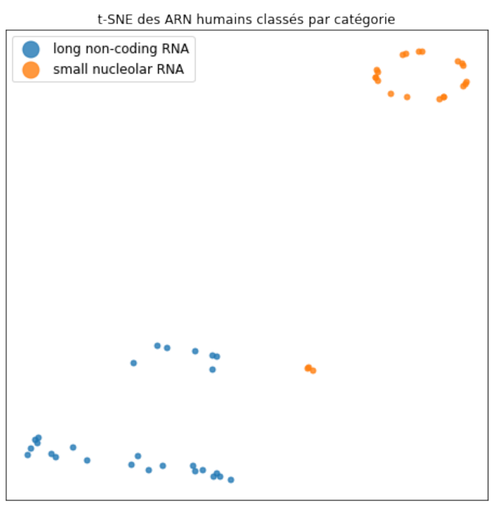


Figure 2. t-SNE comparison of long non-coding and small nucleolar RNA

The first t-SNE plot reveals distinct clusters among RNA categories, indicating that RNA-FM effectively captures meaningful sequence similarities. MicroRNAs and small nucleolar RNAs form tight, well-defined clusters, likely due to their conserved structures, while long non-coding RNAs display a more dispersed pattern, reflecting their functional diversity. It also highlights the distinct properties of human RNA that differentiate it from random RNA sequences.

The second plot, focusing on only two RNA classes, shows clear separability, confirming RNA-FM's ability to differentiate between RNA types.

Deep Learning Architecture

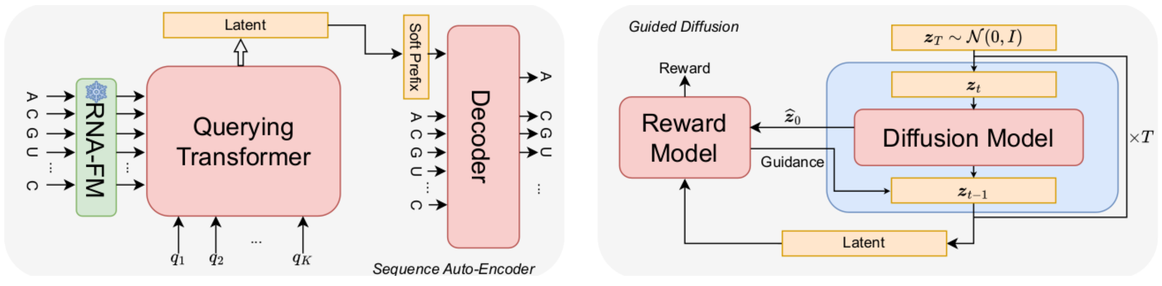


Figure 3. Schematic representation of the architecture, including the embedding by RNA FM, the summarization of the latent space with a Transformer, and the reward-guided diffusion model.

Reward guided diffusion : how to create new sequences that are stabler

A transformer is used to reduce the dimension of the latent space, to sum it up. It provides a vector of fixed dimension K whatever the sequences' number of tokens.

$$\mathcal{L}_{\text{reconstruct}}(\psi, \phi; x) = -\log p_{\text{Dec}_\psi}(x_1, x_2, \ldots, x_L \mid \text{QFormer}_\phi(\text{Enc}(x)))$$

A diffusion model is then built on this summarized latent space. To do so a denoising network is trained to predict the added gaussian noise at each step. The denoiser is a transformer. The architecture eliminates the need for truncation or padding in discrete diffusion.

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0=\text{QFormer}_\phi(\text{Enc}(x))}\mathbb{E}_{t,\epsilon}\left\|\epsilon - \text{Denoiser}_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2$$

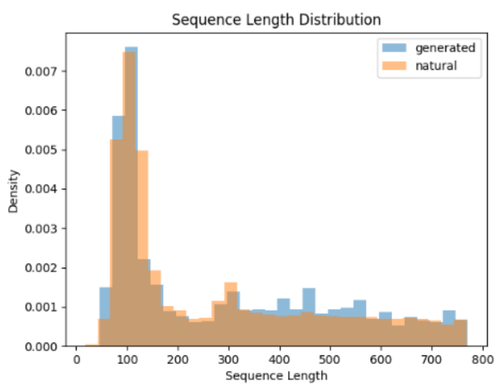
The guided generation block enables to generate RNA sequences of a desired greater stability. The dataset needs to be labeled, (x, r) where r is the measured stability. A network is trained on the latent space to predict the reward.

$$\mathcal{L}(\xi) = \mathbb{E}_{(x,r)}\ell(R_\xi(z), r)$$

This guides denoising toward higher reward zones by estimating the clean sample and using the gradient of the reward loss with respect to the target stability.

$$\text{predicted noise} = \text{Denoiser}_\theta(z_t, t) + \sqrt{1-\bar{\alpha}_t} \cdot \lambda \cdot \nabla_{z_t}\ell(r^*, R_\xi(\hat{z}_0))$$

Generating sequences



In order to assess the quality of generated sequences, different metrics are computed between each generated sequence and the base dataset. This determines how plausible they are. The length of the generated sequences is also compared to the original dataset.

Figure 4. Distributions of both generated and real RNA sequences

- **Minimum Levenshtein Distance:** The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one sequence into another. $d_{\text{min}}(x_{\text{gen}}, \mathcal{R}) = \min_{x_{\text{ref}}\in\mathcal{R}} d(x_{\text{gen}}, x_{\text{ref}})$
- **Minimum 4-mer Distance:** A 4-mer is a subsequence of 4 consecutive nucleotides in an RNA sequence. The 4-mer frequency vector $v(x)$ for a sequence x contains the normalized counts of all possible 4-mers. The minimum 4-mer distance between a generated sequence x_{gen} and a reference set $\mathcal{R}$ is given by: $d_{\text{4-mer}}(x_{\text{gen}}, \mathcal{R}) = \min_{x_{\text{ref}}\in\mathcal{R}} \|v(x_{\text{gen}}) - v(x_{\text{ref}})\|_2$.

Figure 5. Distribution of both distances between generated, real and random sequences

References

- Mattick, J.S. *The central role of RNA in human development and cognition*, 2011.
- Huang, K., Yang, Y., Fu, K., Chu, Y., Cong, L., Wang, M. *Latent Diffusion Models for Controllable RNA Sequence Generation*, 2024.
- Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., King, I., et al. *Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions*, 2022.
- Li, J., Li, D., Savarese, S., Hoi, S. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*, 2023.
- Gruver, N., Stanton, S., Frey, N. C., Rudner, T. G. J., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., Wilson, A. G. *Protein design with guided discrete diffusion*, 2023.