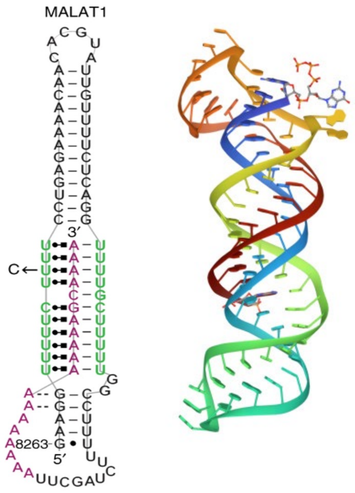
Small Molecule Targeting of MALAT1 Using Generative Artificial Intelligence

Disha Sanwal, Lukas Herron, and Pratyush Tiwary
University of Maryland, College Park

MALAT1 lncRNA – potential drug target

Noncoding RNAs (ncRNAs) are functional RNA molecules that possess key roles in several cellular processes, for example gene regulation and metabolism. Mutations in ncRNAs are associated with numerous human diseases such as cancer, cardiovascular, and neurological disorders [1]. Accordingly, ncRNAs have emerged as potential drug targets for these diseases, making them extremely medically relevant.

The metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) long ncRNA (> 200 nucleotides) is abundantly expressed and elevated MALAT1 expression is correlated to larger tumor size and advanced tumor stage, making it a potential cancer biomarker and target for cancer treatment [2].

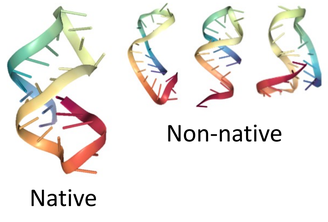


Its unique 3' end triple helix makes it a suitable drug target; helix and pseudoknot form the triple helix. It is usually optimal for a drug molecule to bind to just one target rather than potentially cause off-target drug toxicity. MALAT1's distinctive triple helix makes this easier for the drug.

Figure 1: Secondary and tertiary structures of MALAT 1 [3][4]. Triple helix found at top of tertiary graphic.

RNA structure → determines function

Targeting MALAT1 requires detailed structural information but only a few experimentally solved structures of MALAT1 are known. Since RNAs are flexible by nature, they adopt a broad ensemble of conformation rather than a single native state.



RNAs do not crystallize well and alternate experimental options are costly, requiring the use of novel computational techniques that integrate MD with generative AI to achieve our sequence-to-structural ensemble pipeline.

Sequence-to-structural ensemble pipeline

Molecular dynamics (MD), which simulates chemical systems at the atomic level by numerically integrating the equations of motion, also struggles to understand RNA structure and small molecule recognition due to highly complex RNA dynamics occurring over long timescales.

Replica Exchange Molecular Dynamics (REMD) has previously been implemented to accelerate RNA sampling [5]. However, REMD scales poorly with system size, making the technique inapplicable to MALAT1.

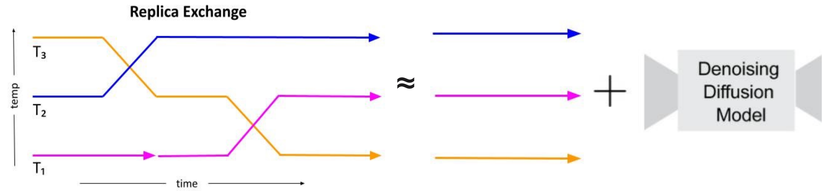


Figure 2: Replica exchange is replaced with independent MD simulations and DDPM to accommodate MALAT1's large size.

Rosetta [6] is used to generate initial folded structures of MALAT1 from its sequence, which is then used to seed independent, unbiased MD simulations at different temperatures. The data that comes out needs to be combined in a physically realistic way. Usually FEP calculations are used, but they are too expensive.

Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Models (DDPMs) – generative AI models that transform noise to data – are used to combine the data from the MD simulations and learn about RNA structure across temperatures [7]. By doing this over a wide range of temperatures, then generating samples at the physiological temperature 310 K, many diverse conformations to be identified and checked for thermodynamic consistency. This MD-AI combination yields an enriched structural ensemble at hard-to-sample physiological temperatures.

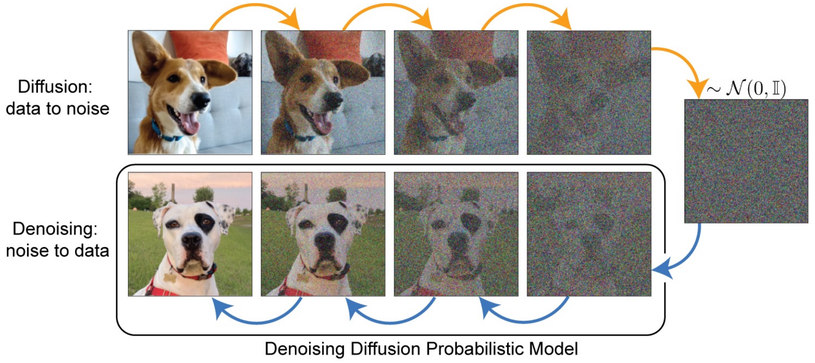


Figure 3: The DDPM process, transforming between data and noise. As seen, new and realistic samples are generated from input data provided.

Preliminary simulation results

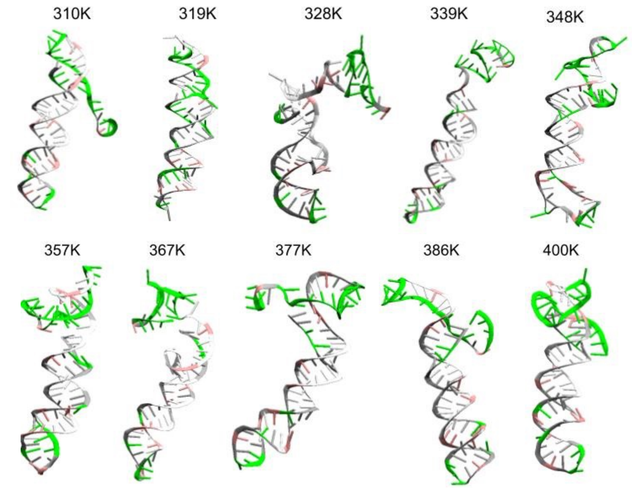


Figure 4: MD simulations over 310-400K from seeded Rosetta starting structures of MALAT1.

As no unfolding is present, higher temperature simulations are needed next. The triple helix is not recovered, but the structures are reasonably accurate when comparing to the secondary structure.

Post-DDPM, small molecule targeting of MALAT1 will be investigated by docking, followed by increasingly rigorous FEP methods to estimate accurate binding free energies and rule out false positives. The binding free energies will be assessed to determine if any small molecule emerges as a potential ligand and sent to NCI for experimental verification.

Acknowledgements

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM142719. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors also thank Deepthought2 and XSEDE (project CHE180027P) for the computational resources used in this work.

References

[1] Lekka, E. & Hall, J. Noncoding RNAs in disease. *FEBS Lett.*, **2018**, *592*, 2884-2900.
[2] Shi, X.S.; Li, J.; Yang, R.H. *et al.* Correlation of increased MALAT1 expression with pathological features and prognosis in cancer patients: a meta-analysis. *Genet Mol Res.*, **2015**, *14(4)*, 18808–19.
[3] Brown, J., Bulkley, D., Wang, J. *et al.* Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nat Struct Mol Biol,* **2014**, *21*, 633–640.
[4] RCSB PDB. 4PLX: Crystal structure of the triple-helical stability element at the 3' end of MALAT1. Rcsb.org, **2014**
[5] J. Earl and M. W. Deem, Parallel Tempering: Theory, Applications, and New Perspectives, Phys. Chem. Chem. Phys., **2005**, *7*, 3910.
[6] A. M. Watkins, R. Rangan, and R. Das, FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds, **2018**, Structure 28, 963.
[7] Wang, Y.; Herron, L.; Tiwary, P. From data to noise to data for mixing physics across temperatures with generative artificial intelligence. *Proc Natl Acad Sci USA*, **2022**, *119*(32) e2203656119.