## ABSTRACT

Computational and Experimental Investigations of
Noncoding RNA Structures

Zion Ruth Perry

2026

Noncoding RNAs (ncRNAs) are central to regulating gene expression and many aspects of cellular function. Diverse classes of ncRNAs have been identified, but the molecular mechanisms underlying their functions often remain unclear. RNA structure provides a valuable starting point for investigating these mechanisms, since structure can influence intermolecular interactions, RNA stability, and modulate various aspects of activity within cells. Understanding RNA structure is also important from a fundamental perspective to uncover the biophysical principles governing macromolecular folding. Additionally, dysregulation of ncRNAs has been linked to numerous human diseases, including cancer. Determining their structures can provide mechanistic insights into these diseases and reveal new therapeutic targets.

Computational and experimental approaches are complementary strategies for determining RNA structure. Experimental structural data can guide computational folding algorithms to generate more accurate structure models. Quantitative and bioinformatic analyses based on these predictions can help identify discrete or conserved structural elements that may have functional significance. These structures can then be used to generate mechanistic hypotheses for experimental testing. The central goal of this dissertation is to both determine and develop tools for modeling RNA structure that will enable better investigation of RNA function.

One class of noncoding RNAs is the long noncoding RNAs (lncRNAs), which are large transcripts with diverse regulatory functions. However, few structures of lncRNAs have been experimentally determined, particularly in living cells. Secondary structure, the patterns of base-

pairing that define higher-order RNA structure, is a tractable starting point for understanding lncRNA structure. **Chapter 2** therefore presents the secondary structure of the functionally relevant lncRNA HOTAIR. Because HOTAIR has been implicated in promoting breast cancer progression, its structure was determined in a metastatic breast cancer cell line. The SHAPE-MaP chemical probing technique was used to determine the global architecture of HOTAIR and identify discrete structured modules. This was compared to a corresponding structure model solved under *in vitro* conditions to identify structures unique to the cellular context in cancer. Additionally, multiple sequence alignments across primates reveal conservation patterns over evolutionary time and covariation within specific domains of HOTAIR. These results demonstrate the utility of chemical probing for determining lncRNA structure in cells, which when coupled with quantitative and bioinformatic analyses, can inform mechanistic hypotheses.

Solving the experimental three-dimensional structures of RNAs is technically challenging but more feasible for smaller, structurally compact ncRNAs. **Chapter 3** thus evaluates a newly developed reconstruction program Arena for generating full-atomic RNA structure models from coarse-grained structures. Arena was benchmarked on a dataset of experimentally determined ncRNA structures, and various quality control metrics were used to assess the accuracy of the predictions. Starting from a coarse-grained RNA structure, Arena first fills in missing non-hydrogen atoms and then assigns secondary structure features. Arena then performs iterative refinement of bond lengths and angles, corrects base and base pair conformations, and removes steric clashes to generate a full-atomic model. Arena is a user-friendly, open-source tool for reconstructing and refining ncRNA structures, and the generated models serve as useful starting points for mechanistic studies or molecular simulations.

Together, this work integrates computational and experimental approaches to examine RNA structure at multiple levels: primary sequence conservation and secondary structure of a lncRNA in living cells and three-dimensional reconstruction of diverse ncRNAs. This work shows that RNA folds into intricate structures that provide a foundation for testing their functional mechanisms. This work both advances our understanding of RNA structure and may be useful for developing RNA-based technologies and therapies where controlling or designing RNA structures would be useful for achieving desired functions.

Computational and Experimental Investigations of

Noncoding RNA Structures

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Zion Ruth Perry

Dissertation Director: Dr. Anna Marie Pyle

May 2026

**TABLE OF CONTENTS**

This dissertation is dedicated to my parents, for nurturing in me

a sense of wonder for the natural world.

*"If I had influence with the good fairy who is supposed to preside over the christening of all children, I should ask that her gift to each child in the world be a sense of wonder so indestructible that it would last throughout life."*

–Rachel Carson, *The Sense of Wonder*

**ACKNOWLEDGEMENTS**

First and foremost, I would like to thank my advisor, Dr. Anna Marie Pyle, for wholeheartedly supporting me for the past five years. Anna accepted me into her lab, though I had no previous experience working with RNA, and gave me the freedom to pursue my academic goals of combining experimental and computational work. Her feedback was always timely and incisive, pushing me to think more deeply about my work.

I would like to thank my dissertation committee members, Dr. Matthew Simon, Dr. Karla Neugebauer, and Dr. Sigrid Nachtergaele. I came away from every committee meeting with renewed clarity on project directions. Their creative insights and scientific rigor enabled this work to come to fruition.

I would like to thank all current and former Pyle lab members, who created a scientifically stimulating and collaborative atmosphere. What struck me the most was how eager anyone was to help when I had a question or needed to learn a new protocol. Special gratitude goes to my mentors, Dr. Ananth Kumar, who guided me on the work presented in **Chapter 2**, and Dr. Chengxin Zhang, who guided me on the work presented in **Chapter 3**. Thank you to my lab mates who joined the lab right before me, Michelle Luo and Tanja Hann, for sitting down with me to go through data and becoming friends to journey through this process together.

I would like to thank my Yale BQBS cohort and the entire Molecular Biophysics and Biochemistry (MB&B) department for being a source of encouragement to me, especially during my first year of graduate school. I am also thankful to have been a part of other Yale communities that enriched my graduate school experience, including the McDougal Graduate Teaching Fellows and Physical and Engineering Biology programs.

I would like to express my sincere appreciation to my previous scientific mentors who provided formative undergraduate research experiences. Thank you to Dr. Tania Baker, in whose lab I learned about protein biochemistry under the mentorship of Dr. Sora Kim, and to Dr. Tanja Kortemme, in whose lab I learned about computational protein design under the mentorship of Dr. Anum Glasgow. In both experiences, I was new to each field, and my mentors displayed utmost patience and kindness in teaching me. Though I switched to the 'dark side' of RNA for graduate school, the skills I learned from them have helped carry me through my graduate school journey.

I would like to thank the various church communities, mentors, and friends that have supported and prayed for me throughout graduate school. I am thankful to not be running the race of life alone. Every time I wanted to give up, someone was there to give a timely word. I especially thank my parents, brother, and grandmother, who have believed in me every step of the way and my dad for encouraging me to apply to Yale. Finally, thank you to N for cheering me on to the finish line and being eager to learn all about a molecule called HOTAIR.

I am in awe that one could spend multiple lifetimes studying one molecule and still scarcely understand its complexity and myriad interactions. I am thankful to be a tiny part of the worthy endeavor of understanding the molecules of life.

## LIST OF FIGURES

## LIST OF TABLES

# 1. CHAPTER 1: INTRODUCTION

## 1.1 THE MARVELOUS NONCODING GENOME

*Among all the marvels that I have discovered in nature, these are the most marvelous of all.*
–Antonie van Leeuwenhoek

The beauty of life we see around us is partially a result of the millions of biological macromolecules in a single cell and the interactions between them. During the advent of microscopy, scientists were able to visualize single cells, such as those referred to as 'animalcules' by Antonie van Leeuwenhoek in his pioneering observations [1]. Today, structural biology techniques enable visualization of single molecules at near-atomic resolution [2]. In parallel, advances in supercomputing allow scientists to run computational simulations of billions of atoms at an unprecedented scale. Despite the acceleration of innovation over the past several centuries and our ability to 'see' molecules, our understanding of their precise mechanisms remains largely incomplete, and new scientific discoveries continue to challenge longstanding biological paradigms.

One foundational paradigm is the central dogma of molecular biology, which describes the flow of genetic information in the cell. In this model, DNA is transcribed into messenger RNA (mRNA), which in turn is translated into proteins that carry out the many biological functions in the cell [3]. Although this model is an instructive starting point, many scientific discoveries have since revealed the seemingly limitless complexity of gene regulation. In particular, the 'centrality of RNA' has become increasingly recognized, opening up new areas of scientific investigation [4]. Distinct from its role as an intermediate messenger for protein translation, RNA boasts an incredible diversity of 'noncoding' functions and is essential for regulating many cellular processes.

Current gene annotations estimate that protein coding genes comprise about 25% of annotated human genes (GENCODE Release 49). Although proteins are generally viewed as the 'workhorses' of the cell, the greater portion of the genome is transcribed into various noncoding RNAs (ncRNAs) that play crucial regulatory roles in the cell. Large proportions of noncoding genes are more prevalent across mammalian genomes than bacteria, for example, and likely contribute to increased organismal complexity [5]. The best-characterized ncRNAs include ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), which are central in protein synthesis, but numerous other classes of ncRNAs—snRNAs, snoRNAs, miRNAs, piRNAs, and circRNAs, to name a few—contribute vital roles.

One important class of ncRNAs are the long noncoding RNAs (lncRNAs), loosely defined as transcripts longer than 200 nt, although a more recent consensus increased the threshold to 500 nt to better distinguish them from other RNA classes [6]. Similar to mRNAs, many lncRNAs are transcribed by RNA polymerase II and undergo capping, polyadenylation, and splicing. While useful for general classification, there are many exceptions to these definitions, underscoring the diversity of these long transcripts. There are upwards of 35,000 annotated lncRNA genes in the human genome (GENCODE Release 49), but most of them have unknown functional roles. Distinguishing functionally viable lncRNAs from transcriptional 'noise' remains a challenging but important task.

For those that have been studied, lncRNAs exhibit diverse functional mechanisms, and have been noted for their roles in transcriptional and epigenetic regulation. In epigenetic regulation, lncRNAs can activate or repress gene expression near their original site of transcription *in cis* or at distal loci *in trans* [7]. In some cases, the effect of a lncRNA occurs at its DNA locus, which may function as a promoter, recruit transcription factors, or regulate alternative splicing of

neighboring genes, as shown for the lncRNAs Bendr and Blustr [8]. Additionally, some isoforms of lncRNAs encode short peptides or proteins, such as the lncRNA SRA and its corresponding protein product SRAP [9]. Because these different roles at the DNA, RNA, and protein levels are not mutually exclusive, dissecting their functional mechanisms remains challenging.

LncRNA loci are often organizationally complex. A lncRNA gene may be positioned in an intergenic or intronic region, be divergent to neighboring genes, or overlap neighboring genes in an antisense orientation [10]. LncRNAs also can form multiple transcripts through alternative transcription start sites (e.g., Tsix [11]), alternative cleavage and polyadenylation (e.g., CCAT1-L [12]), and alternative splicing (e.g., Neat1 [13]). The fact that noncoding elements are not constrained by the genetic code like protein coding genes likely contributes to this diversity and helps explain their rapid phylogenetic divergence across mammalian genomes [6]. Despite exhibiting low sequence conservation across mammals, many lncRNAs participate in essential RNA-RNA or RNA-protein interactions, and dysregulation of these interactions can lead to disease (e.g., NORAD [14]). LncRNAs are deeply embedded within cellular interaction networks, but the details of how their interactions are specified and regulated remain unclear. One way of addressing this challenge is by understanding the role of RNA structure in regulating lncRNA function.

## 1.2 The Intricacies of RNA Structure

The principle that structure determines function applies across biological scales, from organisms to individual molecules. At the molecular level, RNA structure is crucial for regulating and controlling aspects of its functional mechanism. RNA architecture is commonly categorized into four hierarchical levels: the primary, secondary, tertiary, and quaternary levels. The primary level corresponds to the linear nucleotide sequence of adenine (A), cytosine (C), uracil (U), and

guanine (G) residues. Secondary structure describes the base-pairing interactions that allow RNA to form either helical or single-stranded regions. Tertiary structure describes the long-range interactions and three-dimensional fold of an RNA molecule that gives it its functional conformation in cells. Finally, quaternary organization describes the intermolecular interactions that influence RNA structure and function. Additional layers of structural regulation include RNA modifications (which can change base pairing properties), alternatively spliced isoforms (different sequences may fold into different structures), and subcellular localization (which may subject RNA to different binding partners, which in turn may modulate structure), which further diversify RNA function.

Although classical diagrams of RNA secondary structures look like networks of 'ladders, loops, and bubbles' [15], closer inspection reveals that RNA can participate in both canonical Watson-Crick and many non-canonical base-pairing interactions that enable more structural diversity than that seen in DNA. Common features of RNA secondary structures include hairpin stem loops, multi-loop junctions, and helical stacks, while long-range or tertiary level interactions include pseudoknots and kissing loop interactions [16]. Functional RNA secondary and tertiary structures have been found across many classes of ncRNAs, including the cloverleaf and non-canonical shapes of tRNAs [17] and pseudoknots in a viral RNA genome [18, 19].

Because RNA structure is closely connected to function, determining lncRNA structures could provide insight into their functional mechanisms. However, obtaining high-resolution tertiary structures using X-ray crystallography or cryo-electron microscopy, which would provide the greatest level of structural detail, is technically challenging for such large RNAs. In 2025, only 9% of structures deposited in the Protein Data Bank (PDB) were RNA structures (PDB Statistics), and many correspond to structurally compact RNAs such as rRNAs and introns. To date, no high-

resolution 3D structure has been determined for a lncRNA. Methods to determine secondary structure, which have become well-established for RNA, therefore offer a more experimentally tractable starting point and can be used to find discrete structural modules in a long RNA to guide mechanistic studies. However, most lncRNA secondary structures have been determined under controlled conditions *in vitro*, whereas *in cellulo* structures would be more valuable for understanding how binding interactions, macromolecular crowding, and intracellular ionic conditions influence the functional conformations of an RNA.

Chemical probing techniques have been developed as a rapid and indirect method of assessing RNA secondary structure. In probing experiments, chemical probes modify nucleotides in a way that reflects their structural state. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) reagents, for example, react with the ribose 2'OH on all four nucleotides, providing a read-out of local backbone flexibility [20, 21]. Paired nucleotides are typically more constrained and thus show lower SHAPE reactivity, while single-stranded regions tend to be more flexible and thus show higher SHAPE reactivity. Multiple SHAPE reagents have been developed, including 2A3, which demonstrates improved cell permeability and accuracy for *in cellulo* probing experiments (**Figure 1.1**) [22].

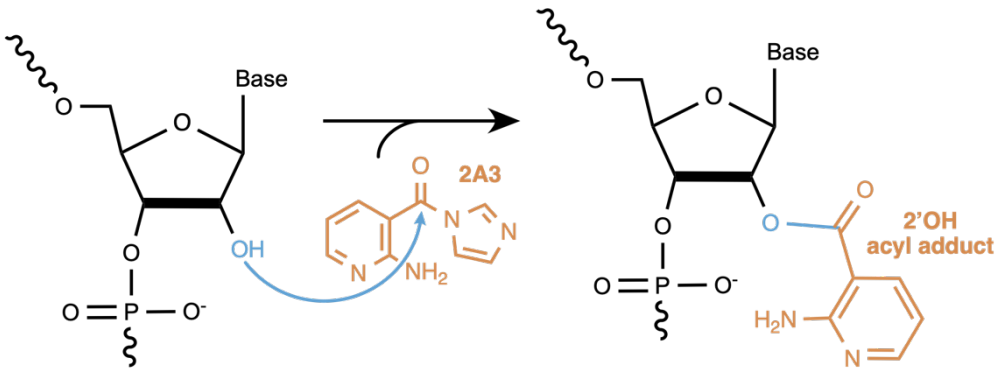



**Figure 1.1 Chemical reaction of the SHAPE reagent 2A3 with RNA.**

The 2'OH on the ribose sugar of a nucleotide residue engages in nucleophilic attack of the SHAPE reagent 2-aminopyridine-3-carboxylic acid imidazolide (2A3) to form a 2'OH acyl adduct. Image created in ChemDraw. Based on Marinus et al. [22].

Orthogonal probing approaches, such as dimethyl sulfate (DMS)-based methods also report on structural accessibility but modify the Watson-Crick faces, thus providing a more direct readout of base-pairing status [23]. In the cellular context, protein binders and tertiary folding may also influence probe reactivity. A major advance is that chemical probing approaches can be combined with mutational profiling, where chemical adducts are incorporated as mutations into complementary DNA (cDNA) during reverse transcription of the modified RNA. Mutations are then counted with high-throughput sequencing, rather than previous gel-based readouts, and used to generate per-nucleotide SHAPE reactivity profiles. These profiles can be used as input for computational RNA secondary structure prediction programs to improve predictions by combining sequence-based minimum free energy constraints with the experimental data. Additionally, other chemical probing techniques have been developed to interrogate information about higher-order or tertiary interactions (e.g., Tb-seq [24] and RING-MaP [25]). These approaches now readily allow determination of RNA structures in cells as a crucial first step to mechanistic studies.

## 1.3 The Long Noncoding RNA HOTAIR

The HOX transcript antisense intergenic RNA (HOTAIR) is a long noncoding RNA transcribed from the *HOXC* gene locus and was first discovered in 2007 [26]. Its genomic location within a *HOX* cluster is significant given the established role of *HOX* genes in early organismal development. HOTAIR is highly expressed during development in certain cell types (i.e., fibroblasts), but its aberrant overexpression in adult tissues has been associated with the progression of a variety of cancers [27-29]. Notably, HOTAIR was found to be overexpressed in

patient-derived metastatic breast cancer samples compared to primary tumor samples and normal breast epithelial cells [30]. Further experiments showed that forced overexpression of HOTAIR in mouse models promoted breast cancer metastasis into the lung.

The proposed mechanism by which HOTAIR promotes tumor progression involves functioning as a scaffold for chromatin modifying complexes, such as PRC2 and LSD1, to direct the epigenetic silencing of tumor suppressor genes [31]. Overexpression of HOTAIR has been shown to change the distribution of LSD1, for example, which enables the epithelial-to-mesenchymal transition, a critical step in cancer metastasis [32, 33].

Multiple isoforms have been annotated in the UCSC Genome Browser (**Figure 1.2**) many of which have not been functionally characterized. The most variable exons seem to correspond to regions that overlap with the reported 5' and 3' end protein binding sites of PRC2 and LSD1, which suggests the potential for isoform-specific interactions and functions [31, 34]. This work focuses on characterizing the structure of a single isoform present in breast cancer cells, which contains both the PRC2 and LSD1 binding sites.

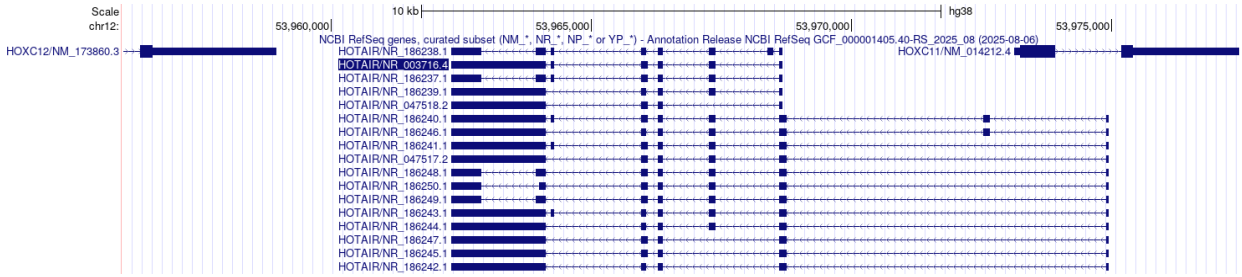



**Figure 1.2 Multiple variants of HOTAIR on the *HOXC* locus of chromosome 12.**
Snapshot from the UCSC Genome Browser (Human assembly GRCh38/hg38) with the NCBI RefSeq annotations shown. Dark blue boxes indicate exons, lines indicate introns, and arrows indicate the direction of transcription. The transcript referenced in this study is highlighted (NR_003716.4).

Previous studies have used *in vitro* transcribed HOTAIR to determine structure models of HOTAIR. Previous work demonstrated that HOTAIR folds into a multidomain secondary structure *in vitro* [35]. Another study proposed a low-resolution three-dimensional model of HOTAIR using atomic force microscopy [36]. However, no structure has been determined in a relevant cellular context. We therefore chose a metastatic breast cancer cell line to determine its secondary structure, which provides a foundation for connecting structural features to functional mechanisms in cancer.

HOTAIR has been identified as a protein binding platform for diverse proteins, indicating it may have other functions beyond epigenetic silencing [37]. Structure may be a useful starting point to understand HOTAIR-protein recognition. RNA modifications are another layer of structural and protein binding regulation. For example, HOTAIR contains an N6-methyladenosine (m6A) modification that contributes to breast cancer invasion [38]. The functional roles of other modifications on HOTAIR, including other m6A sites and 5-methylcytosine (m5C), require further investigation [38-40]. Novel modifications can also be predicted in parallel to structure probing pipelines, providing richer information from the same *in cellulo* samples [41].

Comparative sequence analysis across species can provide additional insight into which RNA structural elements are conserved and therefore likely to be functionally important. Multiple sequence alignments of HOTAIR across broad mammalian genomes have demonstrated low levels of conservation but clear evidence of covariation, in which nucleotide variation preserves base pairing [35, 42]. However, there has been no focused analysis on primate genomes, which may exhibit higher levels of conservation; this is now possible given the recent availability of sequenced primate genomes [43]. Identifying structural elements with conservation and covariation support across primates may help identify structures with important regulatory roles.

## 1.4 Computational Approaches for RNA 3D Structure

In addition to secondary structure prediction for RNA, computational approaches can further our understanding of three-dimensional (3D) RNA structures. Experimental 3D structure determination of RNAs is challenging or results in coarse-grained models, proving most difficult for long, dynamic RNAs. Recently, machine learning methods like AlphaFold have excelled at protein structure prediction [44] and have now been adapted for RNA [45]. Although AlphaFold achieves high accuracy for proteins, RNA structure prediction still lags behind, largely due to insufficiently diverse training data of experimentally solved RNA structures [46]. Rather than predicting a structure *de novo*, RNA structure prediction is more reliable when starting from a few experimentally known atoms. This limitation motivated the development of the program Atomic Reconstruction of RNA (Arena), which takes coarse-grained inputs with as little as one atom per nucleotide and builds a full-atomic RNA structure. Arena performs a series of geometric refinements to generate RNA structures with higher accuracy than state-of-the-art programs when benchmarked on a dataset of diverse ncRNAs. More accurate 3D structures can then enable greater mechanistic insight.

## 1.5 Dissertation Overview

This dissertation explores the structure of ncRNAs at the secondary and tertiary levels. Biological structures provide a detailed roadmap for designing mechanistic studies to gain insight on not just *what* molecules do, but *how* they do it. Furthermore, this dissertation represents a unique opportunity to perform both computational and experimental structural biology work. This shows that integrating multiple methodological approaches can accelerate our understanding of RNA structure and its implications in biological function.

**Chapter 2** discusses the secondary structure of the lncRNA HOTAIR in a metastatic breast cancer cell line. This work represents the first *in cellulo* secondary structure of HOTAIR and identifies structural features unique to the cellular context. Bioinformatic alignments reveal conservation of HOTAIR sequence and secondary structure. This work presents the complex structure of a lncRNA in cells and identifies modules that are candidates for future mechanistic studies on how this RNA may regulate gene expression in a cancer context.

**Chapter 3** discusses the development of the software program Arena, which builds full-atomic 3D RNA structures from coarse-grained input. This software uses geometric rearrangements in compact C++ code to determine structures with higher accuracy and faster runtime than state-of-the-art programs. This shows the utility of using geometric constraints to build RNA structures and contributes open-source software to the structural biology community to build full-atomic structures that are useful for mechanistic studies.

## 1.6 References


1. Lane, N. (2015). The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci*. **370**.

2. Subramaniam, S. (2019). The cryo-EM revolution: fueling the next phase. *IUCrJ*. **6**, 1-2.

3. Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol*. **15**, e2003243.

4. Sharp, P. A. (2009). The centrality of RNA. *Cell*. **136**, 577-580.

5. Shabalina, S. A. & Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol*. **5**, 105.

6. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., Mercer, T. R., Moore, K. J., Nakagawa, S., Rinn, J. L., Spector, D. L., Ulitsky, I., Wan, Y., Wilusz, J. E. & Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol*. **24**, 430-447.

7. Rinn, J. L. (2014). lncRNAs: linking RNA to chromatin. *Cold Spring Harb Perspect Biol*. **6**.

8. Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M. & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. **539**, 452-455.

9. Kawashima, H., Takano, H., Sugita, S., Takahara, Y., Sugimura, K. & Nakatani, T. (2003). A novel steroid receptor co-activator protein (SRAP) as an alternative form of steroid receptor RNA-activator gene: expression in prostate cancer cells and enhancement of androgen receptor activity. *Biochem J*. **369**, 163-171.

10. Ziegler, C. & Kretz, M. (2017). The More the Merrier-Complexity in Long Non-Coding RNA Loci. *Front Endocrinol (Lausanne)*. **8**, 90.

11. Sado, T., Wang, Z., Sasaki, H. & Li, E. (2001). Regulation of imprinted X-chromosome inactivation in mice by Tsix. *Development*. **128**, 1275-1286.

12. Xiang, J. F., Yin, Q. F., Chen, T., Zhang, Y., Zhang, X. O., Wu, Z., Zhang, S., Wang, H. B., Ge, J., Lu, X., Yang, L. & Chen, L. L. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res*. **24**, 513-531.

13. Naganuma, T., Nakagawa, S., Tanigawa, A., Sasaki, Y. F., Goshima, N. & Hirose, T. (2012). Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J*. **31**, 4020-4034.

14. Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y. & Mendell, J. T. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell*. **164**, 69-80.

15. Vicens, Q. & Kieft, J. S. (2022). Thoughts on how to think (and talk) about RNA structure. *Proc Natl Acad Sci U S A*. **119**, e2112677119.

16. Kai, Z., Yuting, W., Yulin, L., Jun, L. & Juanjuan, H. (2019). An efficient simulated annealing algorithm for the RNA secondary structure prediction with Pseudoknots. *BMC Genomics*. **20**, 979.

17. Biela, A., Hammermeister, A., Kaczmarczyk, I., Walczak, M., Koziej, L., Lin, T. Y. & Glatt, S. (2023). The diverse structural modes of tRNA binding and recognition. *J Biol Chem*. **299**, 104966.

18. Wan, H., Jang, H., Xu, L., Abdallah, K. S., Gilbert, W. V. & Pyle, A. M. (2026). A ribosome-bound pseudoknot in the HCV coding region stimulates viral growth by tuning viral translation. *Cell Rep*. **45**, 116739.

19. Wan, H., Adams, R. L., Lindenbach, B. D. & Pyle, A. M. (2022). The In Vivo and In Vitro Architecture of the Hepatitis C Virus RNA Genome Uncovers Functional RNA Secondary and Tertiary Structures. *J Virol*. **96**, e0194621.

20. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods*. **11**, 959-965.

21. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. (2015). Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc*. **10**, 1643-1669.

22. Marinus, T., Fessler, A. B., Ogle, C. A. & Incarnato, D. (2021). A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Res*. **49**, e34.

23. Mitchell, D., Cotter, J., Saleem, I. & Mustoe, A. M. (2023). Mutation signature filtering enables high-fidelity RNA structure probing at all four nucleobases with DMS. *Nucleic Acids Res*. **51**, 8744-8757.

24. Patel, S., Sexton, A. N., Strine, M. S., Wilen, C. B., Simon, M. D. & Pyle, A. M. (2023). Systematic detection of tertiary structural modules in large RNAs and RNP interfaces by Tb-seq. *Nat Commun*. **14**, 3426.

25. Homan, P. J., Favorov, O. V., Lavender, C. A., Kursun, O., Ge, X., Busan, S., Dokholyan, N. V. & Weeks, K. M. (2014). Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci U S A*. **111**, 13858-13863.

26. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. & Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. **129**, 1311-1323.

27. Dong, L. & Hui, L. (2016). HOTAIR Promotes Proliferation, Migration, and Invasion of Ovarian Cancer SKOV3 Cells Through Regulating PIK3R3. *Med Sci Monit*. **22**, 325-331.

28. Loewen, G., Jayawickramarajah, J., Zhuo, Y. & Shan, B. (2014). Functions of lncRNA HOTAIR in lung cancer. *J Hematol Oncol*. **7**, 90.

29. Yang, Z., Zhou, L., Wu, L. M., Lai, M. C., Xie, H. Y., Zhang, F. & Zheng, S. S. (2011). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann Surg Oncol*. **18**, 1243-1250.

30. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S. & Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. **464**, 1071-1076.

31. Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E. & Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science*. **329**, 689-693.

32. Jarroux, J., Foretek, D., Bertrand, C., Gabriel, M., Szachnowski, U., Saci, Z., Guo, S., Londono-Vallejo, A., Pinskaya, M. & Morillon, A. (2021). HOTAIR lncRNA promotes epithelial-mesenchymal transition by redistributing LSD1 at regulatory chromatin regions. *EMBO Rep*. **22**, e50193.

33. Padua Alves, C., Fonseca, A. S., Muys, B. R., de Barros, E. L. B. R., Burger, M. C., de Souza, J. E., Valente, V., Zago, M. A. & Silva, W. A., Jr. (2013). Brief report: The lincRNA Hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. *Stem Cells*. **31**, 2827-2832.

34. Potolitsyna, E., Hazell Pickering, S., Tooming-Klunderud, A., Collas, P. & Briand, N. (2022). De novo annotation of lncRNA HOTAIR transcripts by long-read RNA capture-seq reveals a differentiation-driven isoform switch. *BMC Genomics*. **23**, 658.

35. Somarowthu, S., Legiewicz, M., Chillon, I., Marcia, M., Liu, F. & Pyle, A. M. (2015). HOTAIR forms an intricate and modular secondary structure. *Mol Cell*. **58**, 353-361.

36. Spokoini-Stern, R., Stamov, D., Jessel, H., Aharoni, L., Haschke, H., Giron, J., Unger, R., Segal, E., Abu-Horowitz, A. & Bachelet, I. (2020). Visualizing the structure and motion of the long noncoding RNA HOTAIR. *RNA*. **26**, 629-636.

37. Delhaye, L., De Bruycker, E., Volders, P. J., Fijalkowska, D., De Sutter, D., Degroeve, S., Martens, L., Mestdagh, P. & Eyckerman, S. (2022). Orthogonal proteomics methods to unravel the HOTAIR interactome. *Sci Rep*. **12**, 1513.

38. Porman, A. M., Roberts, J. T., Duncan, E. D., Chrupcala, M. L., Levine, A. A., Kennedy, M. A., Williams, M. M., Richer, J. K. & Johnson, A. M. (2022). A single N6-methyladenosine site regulates lncRNA HOTAIR function in breast cancer cells. *PLoS Biol*. **20**, e3001885.

39. Amort, T., Souliere, M. F., Wille, A., Jia, X. Y., Fiegl, H., Worle, H., Micura, R. & Lusser, A. (2013). Long non-coding RNAs as targets for cytosine methylation. *RNA Biol*. **10**, 1003-1008.

40. Garbo, S., Minotti, S., Marocco, F., Quattrocchi, L., Di Silverio, I., Bontempi, G., Strippoli, R., Tripodi, M. & Battistelli, C. (2025). HOTAIR requires epitranscriptomic modification to exert its pivotal epigenetic role in Epithelial to Mesenchymal Transition. *Cell Death Dis*. **16**, 753.

41. Tavares, R. C. A., Mahadeshwar, G., Wan, H. & Pyle, A. M. (2023). MRT-ModSeq - Rapid detection of RNA modifications with MarathonRT. *bioRxiv*.

42. He, S., Liu, S. & Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol*. **11**, 102.

43. Kuderna, L. F. K., Gao, H., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rousselle, M., Silva, F. E., Agueda, L., Blanc, J., Gut, M., de Vries, D., Goodhead, I., Harris, R. A., Raveendran, M., Jensen, A., Chuma, I. S., Horvath, J. E., Hvilsom, C., Juan, D., Frandsen, P., Schraiber, J. G., de Melo, F. R., Bertuol, F., Byrne, H., Sampaio, I., Farias, I., Valsecchi, J., Messias, M., da Silva, M. N. F., Trivedi, M., Rossi, R., Hrbek, T., Andriaholinirina, N., Rabarivola, C. J., Zaramody, A., Jolly, C. J., Phillips-Conroy, J., Wilkerson, G., Abee, C., Simmons, J. H., Fernandez-Duque, E., Kanthaswamy, S., Shiferaw, F., Wu, D., Zhou, L., Shao, Y., Zhang, G., Keyyu, J. D., Knauf, S., Le, M. D., Lizano, E., Merker, S., Navarro, A., Nadler, T., Khor, C. C., Lee, J., Tan, P., Lim, W. K., Kitchener, A. C., Zinner, D., Gut, I., Melin, A. D., Guschanski, K., Schierup, M. H., Beck, R. M. D., Umapathy, G., Roos, C., Boubli, J. P., Rogers, J., Farh, K. K. & Marques Bonet, T. (2023). A global catalog of whole-genome diversity from 233 primate species. *Science*. **380**, 906-913.

44. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583-589.

45. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C. C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Zemgulyte, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Zidek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D. & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. **630**, 493-500.

46. Bernard, C., Postic, G., Ghannay, S. & Tahi, F. (2025). Has AlphaFold3 achieved success for RNA? *Acta Crystallogr D Struct Biol*. **81**, 49-62.

# 2. CHAPTER 2: CONSERVED STRUCTURAL FEATURES OF THE LNCRNA HOTAIR IN BREAST CANCER CELLS

## 2.1 PREFACE

A version of this chapter is currently being prepared for publication with Anish Beeram and Dr. Anna Marie Pyle. The bioinformatics work on primate genomes was done in collaboration with Anish Beeram (*Section 2.4.5*). This work would not have been possible without contributions from others, including Dr. Allison M. Porman Swain and Dr. Aaron M. Johnson for generously gifting the MDA-MB-231 HOTAIR cells that enabled *in cellulo* probing and Dr. Li-Tao Guo for providing the pCARs08 HOTAIR plasmid for *in vitro* work. Many former and current members of the Pyle Lab contributed thoughtful feedback and assistance, including Dr. Rafael Tavares, Dr. Shivali Patel, Dr. Han Wan, Tanja Hann, Michelle Luo, Lucille Tsao, and Jack Sung. A special thanks goes to Dr. Ananth Kumar for discussions and mentorship that helped lay the groundwork for this study. The Yale Center for Research Computing provided guidance and use of the high-performance computing clusters. This work was funded by the Howard Hughes Medical Institute, the National Institutes of Health (1R01HG011868-01 to A.M.P.), and the National Science Foundation Graduate Research Fellowship Program (DGE-2139841 to Z.R.P.).

## 2.2 ABSTRACT

Long noncoding RNAs (lncRNAs) regulate diverse cellular processes and are frequently implicated in disease, but their functional mechanisms often remain elusive. One such lncRNA, the HOX transcript antisense intergenic RNA (HOTAIR), is a ~2.1 kb mammalian transcript whose overexpression promotes invasion and metastasis in breast cancer. However, the

mechanisms by which HOTAIR influences gene regulation in cancer are poorly understood. To approach this problem through a structural lens, we determined the full-length *in cellulo* secondary structure of HOTAIR using chemical probing in a metastatic breast cancer cell line. The resulting structure shows that HOTAIR adopts a multidomain architecture and has local structural features unique to the cellular context. Comparison between *in vitro* and *in cellulo* chemical probing identifies regions of differential accessibility that may indicate context-dependent molecular interactions or folding. Conservation analyses further reveal that HOTAIR is conserved across primates with evidence of structural covariation in specific domains. Together, these results provide a roadmap for future mechanistic studies of structure-function relationships in HOTAIR and its contribution to gene regulation in cancer.

## 2.3 Introduction

The classical view of RNA as a messenger for protein translation represents only part of RNA's complex roles in cells. Noncoding RNAs (ncRNAs) participate in many essential cellular processes, and their mis-regulation is associated with numerous human diseases. Long noncoding RNAs (lncRNAs) are a class of ncRNAs longer than 500 nucleotides that are typically transcribed by RNA polymerase II and often capped, spliced, and polyadenylated [1]. Since the discovery of the first human lncRNA Xist in 1992 [2], studies of lncRNAs have revealed their diverse regulatory mechanisms at all stages of gene expression. LncRNAs may function *in cis* at their locus of transcription, such as through transcriptional interference or enhancer-like activity, or function *in trans* at distant chromatin or cellular locations [3, 4]. In some cases, the underlying DNA locus may function in transcriptional regulation [5] or the lncRNA may encode for short peptides or

proteins [6, 7]. Despite rapid advances in the lncRNA field, the functional mechanisms of most lncRNAs remain poorly understood.

The lncRNA HOX transcript antisense intergenic RNA (HOTAIR) is a prominent example for which we have an incomplete mechanistic understanding. HOTAIR was first discovered by Rinn et al. through microarray sequencing of the *HOX* loci in human fibroblasts, which revealed a transcript antisense to the *HOXC* locus on chromosome 12 that regulates gene expression *in trans* at the *HOXD* locus on chromosome 2 [8]. Knockdown of HOTAIR reduced histone 3 lysine 27 trimethylation (H3K27me3) and decreased occupancy of the Polycomb Repressive Complex 2 (PRC2) subunit Suz12 at *HOXD* promoters, implicating HOTAIR in epigenetic silencing. Subsequent studies by Tsai et al. identified the binding sites for PRC2 near the 5' end of HOTAIR and for another chromatin silencing complex, Lysine-Specific Demethylase 1 (LSD1), near the 3' end [9]. This led to the model of HOTAIR as a molecular scaffold for coordinating chromatin-modifying complexes. However, later studies found that HOTAIR can mediate transcriptional repression independently of PRC2 binding [10], suggesting that recruitment of chromatin modifiers may occur indirectly and after HOTAIR localization to target genes [11]. Additional studies have since reported interactions between HOTAIR and diverse protein partners [12-14], indicating HOTAIR may have multifunctional roles in gene expression. Although HOTAIR clearly participates in epigenetic silencing at some level, the mechanisms of protein recruitment and target gene regulation remain unresolved.

Although HOTAIR plays a normal role in early development [8], its aberrant overexpression has been extensively linked with cancer progression. In breast cancer, HOTAIR is highly overexpressed in metastatic tumors relative to primary tumors and normal tissue [15]. Overexpression of HOTAIR in breast cancer cell lines confirmed its invasive influence in cell

culture and mouse models, notably in MDA-MB-231 cells, a triple negative and highly invasive breast cancer cell line. Although many noncoding RNAs contribute to breast cancer progression pathways [16], HOTAIR serves as a model oncogenic lncRNA for dissecting how RNA-mediated regulatory mechanisms drive tumor invasion and metastasis. Beyond breast cancer, elevated HOTAIR expression and invasive phenotypes have been reported in many other cancers, including liver [17], colorectal [18], lung [19], and ovarian [20] cancers. Breast cancer serves as a model system to study how HOTAIR contributes to regulatory pathways in disease.

RNA structure is central for regulating its function. Although there are no high-resolution tertiary structures of any lncRNA, secondary structure is a tractable starting point for discovering lncRNA structural modules that may aid functional studies or 3D structure determination of RNA-protein complexes. Chemical probing is an indirect method of determining the secondary structure of RNA. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) reagents modify the 2'OH of the ribose sugar in the backbone proportionate to its flexibility, and generally base-paired nucleotides will be less modified than single-stranded nucleotides [21, 22]. This generates a reactivity profile that can be used to improve *in silico* secondary structure prediction algorithms [23].

The first information about HOTAIR secondary structure was obtained on a full-length HOTAIR transcript that was transcribed, natively purified *in vitro*, and then probed using SHAPE probing with capillary electrophoresis [24]. This work suggested that HOTAIR folds into four modular domains, which were proposed to provide platforms for protein interactors such as PRC2 and LSD1. Since then, advances in chemical probing and sequencing techniques have enabled more accurate RNA structure determination by combining chemical probing with mutational profiling and next generation sequencing. This methodology has been further extended to cellular

contexts, and improved probes have been developed [25]. Despite these advances, there are few published *in cellulo* secondary structures of lncRNAs with the exception of Xist, MEG3, MALAT1, and SChLAP1 [26-30], which provide valuable information. Although *in vitro* structures derived from purified and refolded RNA are informative, the *in cellulo* structure is more informative because molecular interactions, chemical modifications, molecular crowding, or other cellular conditions may influence folding. Particularly when carried out in parallel with *in vitro* probing on natively folded transcripts, the method can yield important insights into lncRNA structure and function.

In this study, we determine the first, full-length *in cellulo* secondary structure of HOTAIR in a breast cancer model system using SHAPE-MaP with the probing reagent 2A3 and an MDA-MB-231 HOTAIR overexpression cell line. We also determined a corresponding *in vitro* structure of HOTAIR using SHAPE-MaP with 2A3, enabling direct comparison of structural accessibility and folding in cellular and *in vitro* contexts. The structures reported here can serve as a guide for further mechanistic studies of HOTAIR structure-function relationships. To provide evidence for the functional relevance of HOTAIR secondary structures, we built large-scale alignments of HOTAIR homologs from unannotated primate genomes. We used these to demonstrate conservation over evolutionary divergence time and to identify covariant base pairs within specific structural domains. Together, this study supports the conservation of functional structural elements within HOTAIR and provides a foundation for investigating how RNA structure contributes to HOTAIR function, particularly in cancer.

## 2.4 RESULTS

### *2.4.1 Structure Probing in Breast Cancer Cells*

We sought to determine an *in cellulo* secondary structure model of full-length HOTAIR using the SHAPE-MaP pipeline (**Figure 2.1A-B**). Given HOTAIR's established role in breast cancer progression, we performed chemical probing in a breast cancer model system. Two breast cancer cell lines often used in functional studies of HOTAIR, MDA-MB-231 and MCF7, expressed little to no endogenous HOTAIR (**Figure 2.2A**) so were not suitable for SHAPE-MaP. We therefore used an MDA-MB-231 HOTAIR overexpression line (MDA-MB-231 HOTAIR) [13], which we found to express abundant HOTAIR in early-passage cells at ~2,000 copies per cell (**Figure 2.2B**). This system thus provided sufficient transcript levels for surmounting the challenges of *in cellulo* SHAPE-MaP.

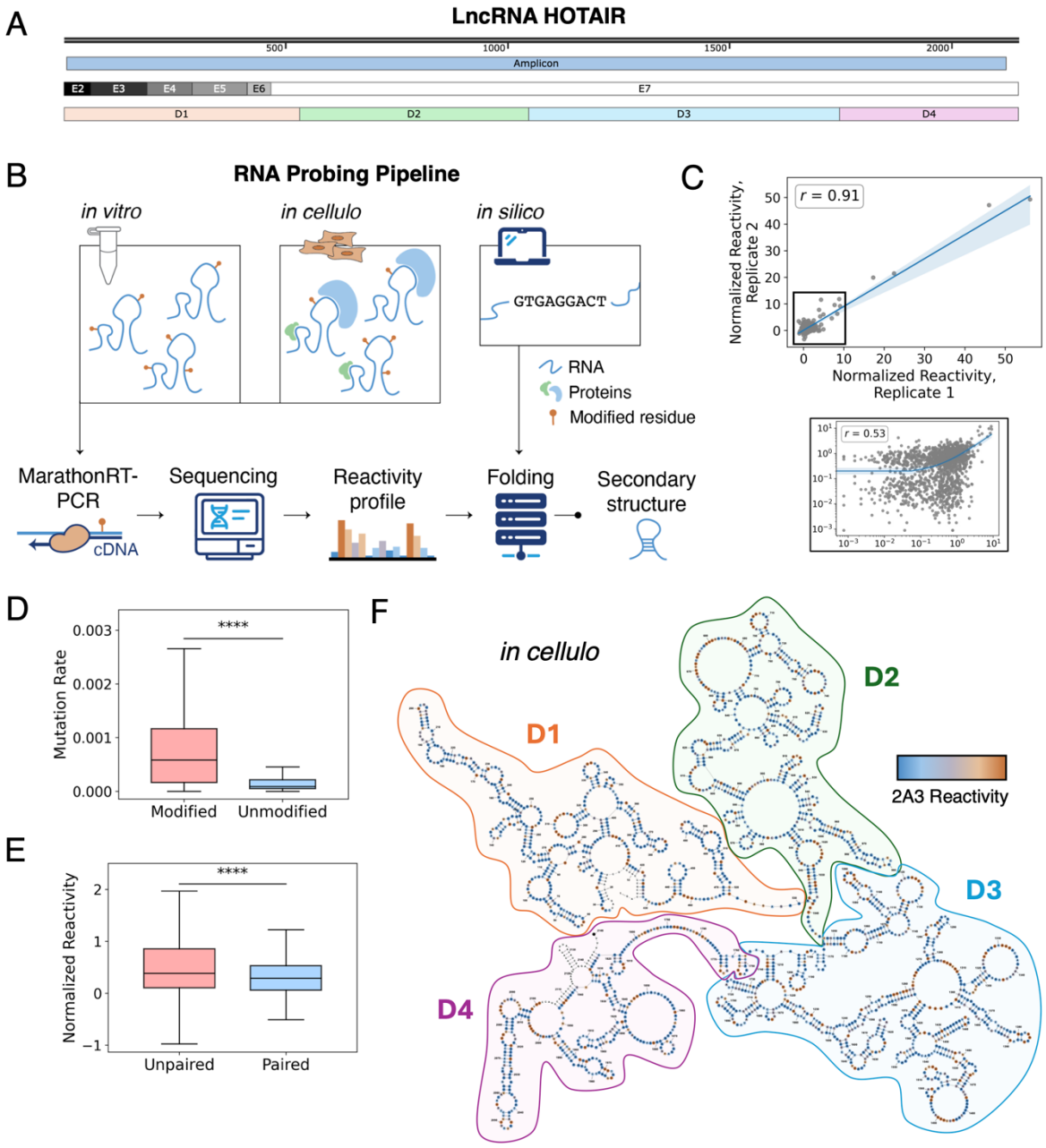



**Figure 2.1 HOTAIR exhibits a complex, multi-domain secondary structure in breast cancer cells.**

(A) Construct of the lncRNA HOTAIR used in this study. Chemical probing data was captured with a single amplicon spanning 7-2122 nt. Exons are designated from exon 2 to exon 7 (E2-E7), and domains of the *in cellulo* secondary structure are designated as Domain 1 to Domain 4 (D1-D4). Image created using SnapGene. (B) RNA chemical probing pipeline for secondary structure prediction across *in vitro*, *in cellulo*, and *in silico* contexts. Image created using Adobe Stock and Illustrator. (C) Global correlation of normalized reactivities between biological replicates of *in cellulo* SHAPE-MaP on HOTAIR for the full data set or for reactivities < 15. The linear regression

line is shown with 95% confidence interval shading. *r*, Pearson's correlation coefficient. (D) Mutation rates of 2A3-modified samples compared to DMSO-unmodified controls for replicate 1 of *in cellulo* SHAPE-MaP on HOTAIR. Visualized with a box-and-whisker plot. (E) Normalized reactivities for residues predicted to be unpaired compared to paired in the secondary structure prediction generated from replicate-averaged *in cellulo* SHAPE-MaP reactivities of HOTAIR. Visualized with a box-and-whisker plot. (F) Secondary structure prediction of HOTAIR using Superfold with replicate-averaged *in cellulo* SHAPE-MaP data. Domains 1-4 are designated as D1-D4. Nucleotides are colored by their normalized reactivity values, orange (≥ 1.0), light orange (≥ 0.75), gray (≥ 0.5), light blue (≥ 0.25), blue ( < 0.25), and white (no data). Structure was visualized with StructureEditor. Statistical significance calculated with two-sided Mann-Whitney U test, **** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, * = $p \leq 0.05$.

Since this cell line had not been used previously for chemical probing, we validated 2A3 probing conditions using 18S rRNA because it is a highly abundant transcript with a well-understood structure. We probed MDA-MB-231 HOTAIR cells with the SHAPE reagent 2A3, which offers enhanced cell-permeability and accuracy compared to other SHAPE reagents [31]. Total cellular RNA was extracted and purified, and a ~1.3 kb region of 18S rRNA was amplified for sequencing and reactivity profile generation via the ShapeMapper2 pipeline [32]. The probing data passed established quality control metrics, including highly correlated reactivity profiles between two biological replicates (Pearson correlation coefficient, r = 0.997; **Figure 2.2C**) and significantly higher mutation rates in 2A3-modified 18S rRNA compared to unmodified controls in both replicates (**Figure 2.2D**).

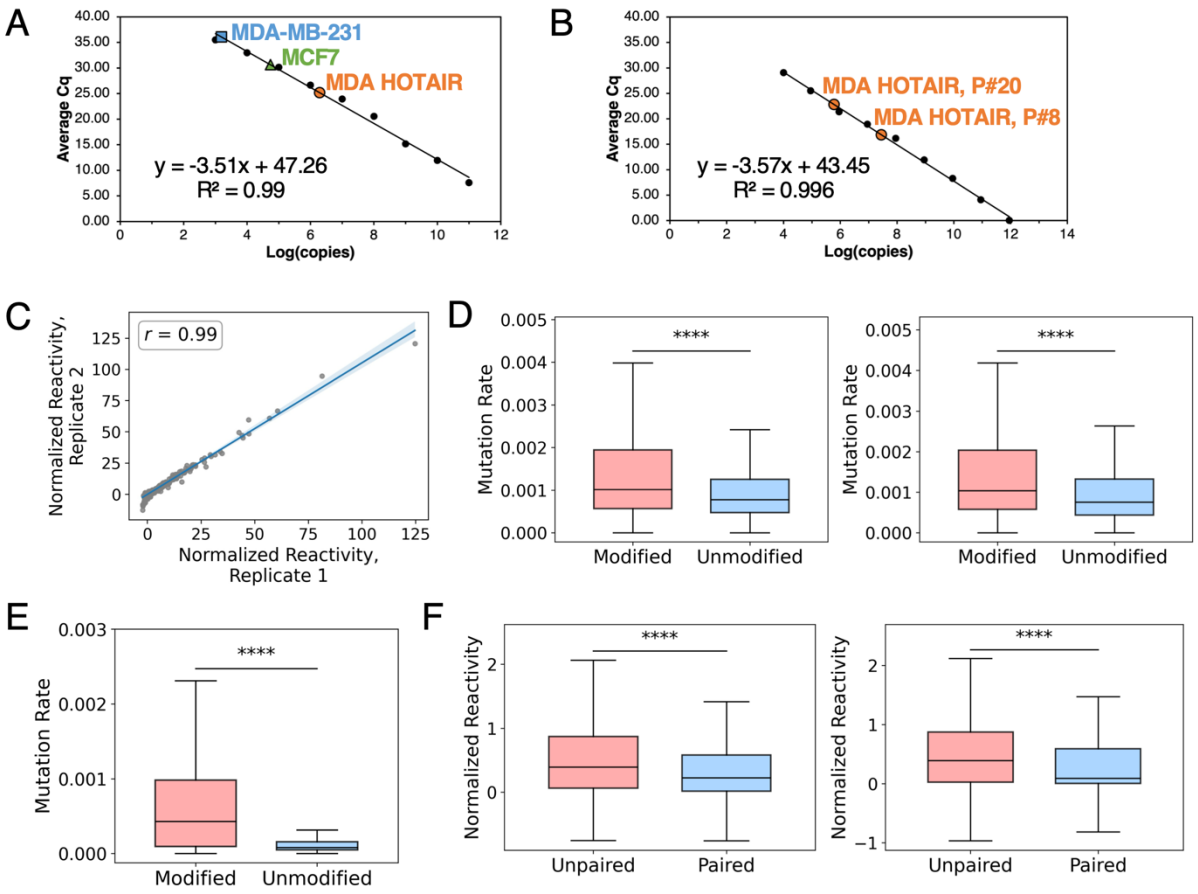



**Figure 2.2 Validation of chemical probing of HOTAIR and 18S rRNA in breast cancer cells.** (A) Expression of HOTAIR in MDA-MB-231, MCF7, and MDA-MB-231 HOTAIR cell lines determined by absolute RT-qPCR. I*n vitro* transcribed HOTAIR was used to generate a standard curve and fitted to a linear regression line to determine copy number. Quantification cycle (Cq) is average of two biological replicates done in duplicate. Plotted in Microsoft Excel. (B) Expression of HOTAIR in MDA-MB-231 HOTAIR lines at two different passage numbers (P#) determined by absolute RT-qPCR. I*n vitro* transcribed HOTAIR was used to generate a standard curve and fitted to a linear regression line to determine copy number. Cq measured in duplicate. Plotted in Microsoft Excel. (C) Global correlation of normalized reactivities between biological replicates of *in cellulo* SHAPE-MaP on 18S rRNA. The linear regression line is shown with 95% confidence interval shading. *r*, Pearson's correlation coefficient. (D) Mutation rates of 2A3-modified samples compared to DMSO-unmodified controls for two replicates of *in cellulo* SHAPE-MaP on 18S rRNA. Visualized with a box-and-whisker plot. (E) Mutation rates of 2A3-modified samples compared to DMSO-unmodified controls for replicate 2 of *in cellulo* SHAPE-MaP on HOTAIR. Visualized with a box-and-whisker plot. (F) Normalized reactivities for residues predicted to be unpaired compared to paired in the secondary structure prediction generated from *in cellulo* SHAPE-MaP reactivities of HOTAIR for replicate 1 (left) and replicate 2 (right). Visualized with a box-and-whisker plot. Statistical significance calculated with two-sided Mann-Whitney U test, **** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, * = $p \leq 0.05$.

After validating chemical probing conditions, the same total cellular RNA was used to amplify a single, ~2.1 kb amplicon spanning 7-2122 nt of HOTAIR. The overexpression construct used in the MDA-MB-231 HOTAIR cells corresponds to the canonical isoform of HOTAIR, containing exons 2 through 7 (**Figure 2.1A**). Reactivity profiles generated using the SHAPE-MaP pipeline showed high global correlation between two biological replicates (r = 0.91; **Figure 2.1C**). Removing the most reactive nucleotides (normalized reactivity > 15) reduced the correlation to weakly positive (r = 0.53). This is consistent with recent *in cellulo* probing studies of lncRNAs, in which lowly reactive nucleotides show increased noise [26]. Mutation rates were significantly higher in 2A3-modified samples compared to unmodified controls in both replicates (**Figure 2.1D and Figure 2.2E**), indicating that the reactivity data was sufficient for proceeding to secondary structure prediction. The MDA-MB-231 HOTAIR cells thus enabled *in cellulo* structural studies of HOTAIR with a single sequencing amplicon.

### *2.4.2 Domain Architecture of HOTAIR*

Reactivity values were averaged from both biological replicates and used to constrain secondary structure predictions using Superfold [23]. As expected, nucleotides predicted to be unpaired showed significantly higher reactivity values compared to nucleotides predicted to be paired (**Figure 2.1E**), indicating that the average structure is consistent with experimental reactivity profiles in addition to the individual replicates (**Figure 2.2F**). The secondary structure model revealed that HOTAIR folds into four complex structural domains in breast cancer cells (**Figure 2.1F**). Domain boundaries were defined as regions comprising complete structures separated by stretches of single-stranded nucleotides and were guided by the prior *in vitro* structure

of HOTAIR (**Table 2.1**). Domain 1 includes exons 2 through 6, while exon 7 spans the end of Domain 1 and extends through Domains 2-4 (**Figure 2.1A**).

**Table 2.1 Domain boundaries for HOTAIR secondary structure models.**

| Sample | Domain 1 (nt) | Domain 2 (nt) | Domain 3 (nt) | Domain 4 (nt) |
|---|---|---|---|---|
| Replicate 1, *in cellulo* | 1-478 | 479-1055 | 1056-1416 | 1417-2148 |
| Replicate 2, *in cellulo* | 1-525 | 526-1123 | 1124-1795 | 1796-2148 |
| Average, *in cellulo* | 1-530 | 531-1046 | 1047-1747 | 1748-2148 |
| Average, *in vitro* | 1-530 | 531-1047 | 1048-1627 | 1628-2148 |
| Prior structure, *in vitro* [24] | 1-530 | 531-1040 | 1041-1513 | 1514-2148 |
| *in silico* | 1-523 | 524-1122 | 1123-1780 | 1781-2148 |

Notably, the prior *in vitro* study reported that HOTAIR folds into four independent structural domains [24]. The nucleotide boundaries of Domains 1 and 2 closely match those reported *in vitro*, whereas Domain 3 is longer and more complex *in cellulo*, with a shorter Domain 4. The conservation of overall domain architecture despite the absence of proteins and other cellular factors *in vitro* suggests that HOTAIR folds into a stable structure that is largely determined by its primary sequence. In contrast, local structural differences within individual domains of the *in cellulo* model may reflect cellular influences that contribute to HOTAIR function.

*2.4.3 Identification of Highly Structured Modules in HOTAIR*

To enable quantitative comparison of *in cellulo* and *in vitro* models, we determined a new *in vitro* secondary structure of HOTAIR using SHAPE-MaP with 2A3. Full-length HOTAIR was *in vitro* transcribed, semi-natively purified by size-exclusion chromatography, and refolded at 25 mM $Mg^{2+}$, as described previously [24]. The probing data passed quality control metrics, including high correlation between biological replicates ($r = 0.99$) and significantly higher mutation rates in the modified compared to unmodified samples (**Figure 2.4A-B**). Comparison of replicate-averaged reactivity profiles from *in cellulo* and *in vitro* experiments show overall agreement, with localized regions of increased or suppressed reactivities that may reflect context-dependent differences in RNA accessibility to the probe (**Figure 2.3A**).

Reactivities from individual replicates and replicate-averaged reactivities were used to constrain secondary structure predictions, with higher reactivities mapping to nucleotides predicted to be unpaired in all cases (**Figure 2.4C-D**). The domain boundaries of the *in vitro* structure determined in this experiment correspond closely to the previous *in vitro* structure. In this experiment, Domain 3 is predicted to be 115 nt longer, which is closer to that observed *in cellulo* (**Table 2.1**). Despite areas of differing accessibility (**Figure 2.3A**), there is substantial shared base-pairing between the *in vitro* and *in cellulo* models, particularly in much of Domains 1 and 2, and the overall four-domain architecture is preserved in both contexts (**Figure 2.3B**).

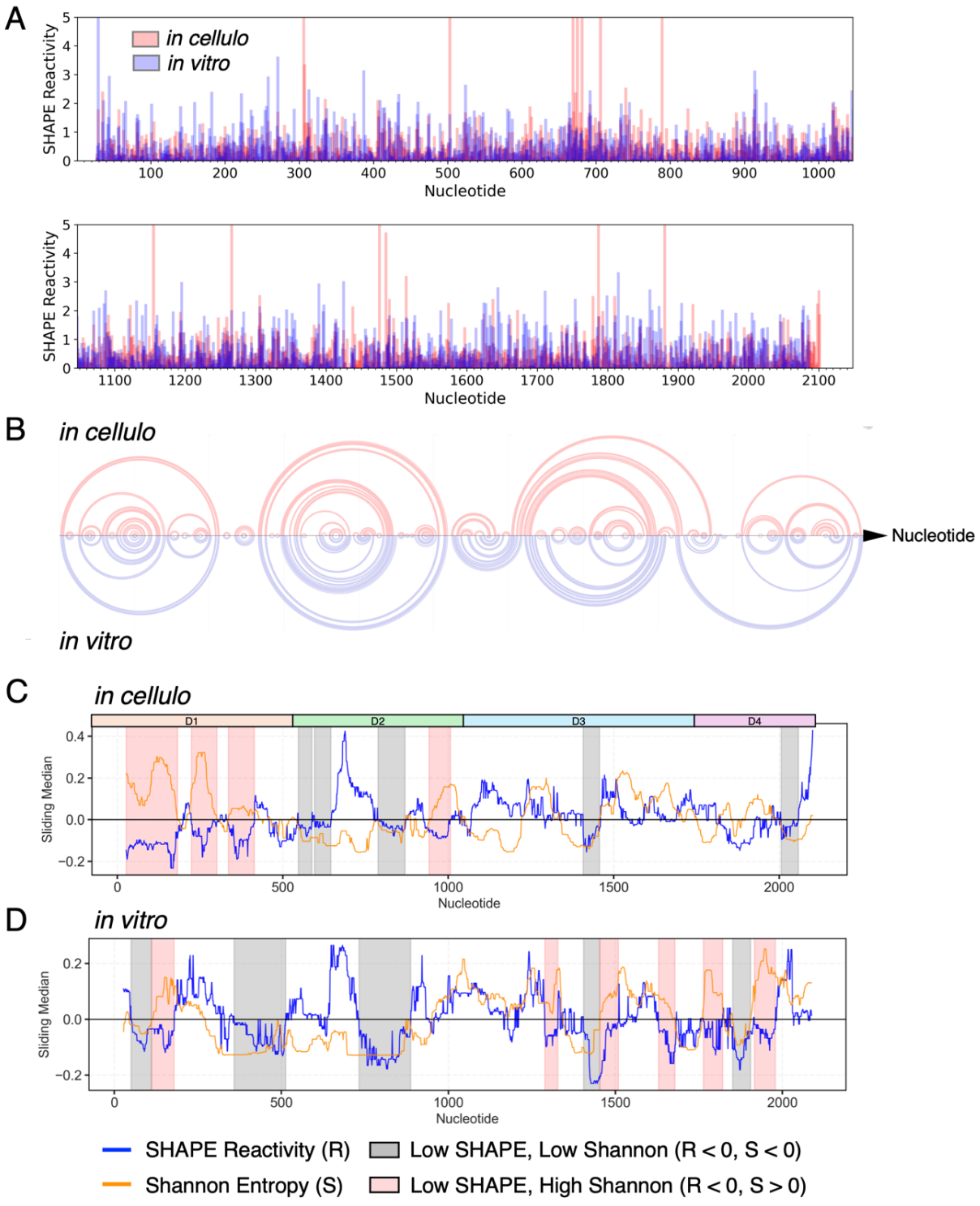



**Figure 2.3 HOTAIR contains highly structured modules across *in cellulo* and *in vitro* contexts.**

(A) Normalized per-nucleotide reactivities from replicate-averaged *in cellulo* (light red) and *in vitro* (light blue) SHAPE-MaP data. The range was limited to < 5 to enable comparison of the

majority of reactivities. (B) Arc plot comparison of the replicate-averaged *in cellulo* and *in vitro* secondary structure predictions. Each arc represents a base pair. Image created with Rchie. (C) Sliding window median analysis of the replicate-averaged reactivity (blue line) and Shannon entropy (orange line) values for *in cellulo* SHAPE-MaP. Domains 1-4 for the *in cellulo* structure are designated as D1-D4. Gray boxes indicate SHAPE reactivity and Shannon entropy below the global median, and light red boxes indicate SHAPE reactivity below the global median and Shannon entropy above the global median. (D) Sliding window median analysis of the replicate-averaged reactivity and Shannon entropy values for *in vitro* SHAPE-MaP. Visualized as in (C).

Reactivity and Shannon entropy criteria are commonly used to evaluate the stability and dynamical characteristics of structured modules within long RNAs. Sliding window median analysis was applied to the replicate-averaged reactivity and entropy profiles (**Figure 2.3C-D**), as well as to individual replicates (**Figure 2.4E-F**). We identified two categories of highly structured regions, defined by low SHAPE reactivity ($R < 0$): those with low Shannon entropy ($S < 0$), suggesting rigid structures, and those with high Shannon entropy ($S > 0$), suggesting conformationally heterogeneous structures.

Comparison of these structural features between *in cellulo* and *in vitro* models reveals domain-level differences in HOTAIR folding (**Figure 2.3C-D**). Domain 1 is highly structured in both contexts yet exhibits higher Shannon entropy across much of the domain *in cellulo* (27-182 nt, 224-301 nt, 336-414 nt) compared to *in vitro*. Domain 2 contains highly structured modules (547-588 nt, 596-645 nt, 788-869 nt) alternating with high reactivity or high Shannon (942-1007 nt) regions *in cellulo*. Both structures have a region of elevated reactivity at ~700 nt, prior to a low SHAPE, low Shannon region, which is also observed in the individual *in cellulo* replicates (**Figure 2.4E-F**).

In contrast, Domains 3 and 4 contain smaller, highly structured and rigid elements *in cellulo* (1408-1457 nt, 2006-2058 nt), whereas the *in vitro* structure exhibits greater conformational heterogeneity in these domains. One structured, low Shannon region is shared

between *in cellulo* (1408-1457 nt) and *in vitro* (1405-1453 nt) contexts. Although the overall domain architecture is similar in both contexts, HOTAIR contains several structured modules unique to the cellular context, which may be important to the cellular roles of HOTAIR.

The experimentally constrained structure models presented here capture structural information beyond a purely thermodynamic *in silico* prediction. Strikingly, both the *in cellulo* and *in vitro* models predict highly similar base pairing in part of Domain 1 (~1-350 nt; Jaccard similarity ~0.8-0.9) that is absent from the *in silico* model (**Figure 2.4G**). In contrast, there are regions of elevated similarity between the experimental and *in silico* models, some of which overlap with low SHAPE, low Shannon regions (**Figure 2.4G, Figure 2.3C-D**). This suggests that HOTAIR contains both thermodynamically stable structures and structures in cells that are different from what is expected thermodynamically.

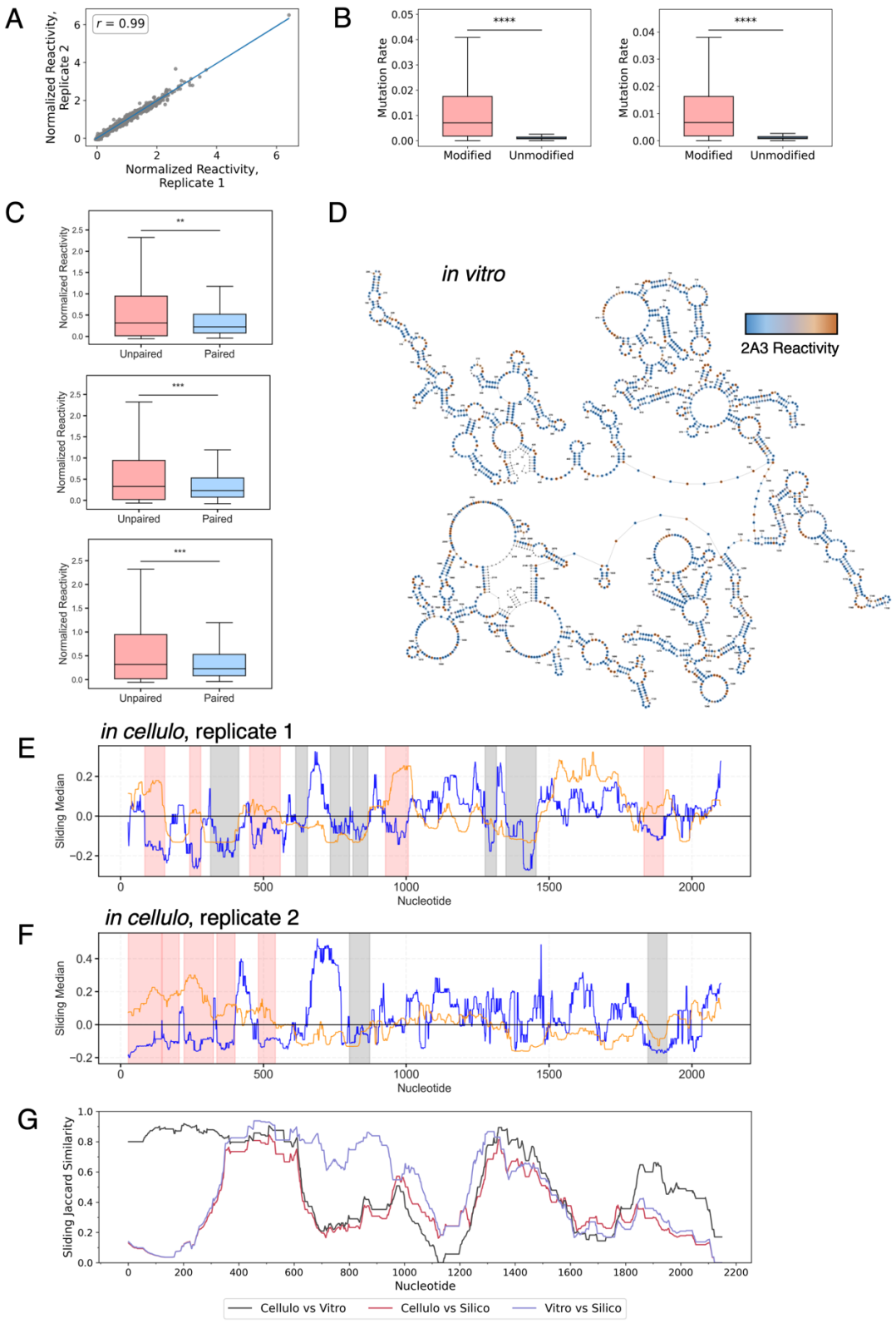



**Figure 2.4 Determination of *in vitro* secondary structure of HOTAIR, structured modules *in cellulo*, and comparison to *in silico* prediction.**

(A) Global correlation of normalized reactivities between biological replicates of *in vitro* SHAPE-MaP on HOTAIR. The linear regression line is shown with 95% confidence interval shading. *r*, Pearson's correlation coefficient. (B) Mutation rates of 2A3-modified samples compared to DMSO-unmodified controls for two replicates of *in vitro* SHAPE-MaP on HOTAIR. Visualized with a box-and-whisker plot. (C) Normalized reactivities for residues predicted to be unpaired compared to paired in the secondary structure prediction generated from *in vitro* SHAPE-MaP reactivities of HOTAIR from replicate 1 (top), replicate 2 (middle), and the replicate-averaged values (bottom). Visualized with a box-and-whisker plot. (D) Secondary structure prediction of HOTAIR using Superfold with replicate-averaged *in vitro* SHAPE-MaP data. Nucleotides are colored by their normalized reactivity values, orange (≥ 1.0), light orange (≥ 0.75), gray (≥ 0.5), light blue (≥ 0.25), blue ( < 0.25), and white (no data). Structure was visualized with StructureEditor. (E) Sliding window median analysis of the reactivity (blue line) and Shannon entropy (orange line) values *in cellulo* SHAPE-MaP, replicate 1. Gray boxes indicate SHAPE reactivity and Shannon entropy below the global median, and light red boxes indicate SHAPE reactivity below the global median and Shannon entropy above the global median. (F) Sliding window median analysis of the reactivity and Shannon entropy values *in cellulo* SHAPE-MaP, replicate 2. Visualized as in (E). (G) Jaccard similarity comparisons in sliding windows for the *in cellulo*, *in vitro*, and *in silico* secondary structure models of HOTAIR. Statistical significance calculated with two-sided Mann-Whitney U test, **** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, * = $p \leq 0.05$.

### *2.4.4 Clusters of Protected and Accessible Sites in HOTAIR*

We next identified nucleotide-level differences in reactivities between replicate-averaged *in cellulo* and *in vitro* reactivities using the deltaSHAPE pipeline [33] (**Figure 2.5A, Figure 2.6**). Positive ΔSHAPE values (> 0) indicate nucleotides that are more protected from modification in cells, which may reflect interactions with cellular factors or structural stabilization. Negative ΔSHAPE values (< 0) indicate nucleotides that are more accessible in cells, which may reflect sites of structural flexibility or unfolding.

Domain 1 contains eleven sites of increased protection and two sites of increased accessibility, each spanning 3-4 nt. The beginning of Domain 1 (16-283 nt) contains a series of highly structured stems that harbor 9 protected sites, which overlaps with the previously determined PRC2 minimal binding region (1-300 nt [9] or 212-300 nt [34]) (**Figure 2.5B**). Many of these sites are predicted to be single-stranded, suggesting they would remain structurally

accessible for protein interactions. This region also displays relatively high Shannon entropy (**Figure 2.3C**), although it is unclear if protein binding dynamics are the source of this predicted conformational heterogeneity.

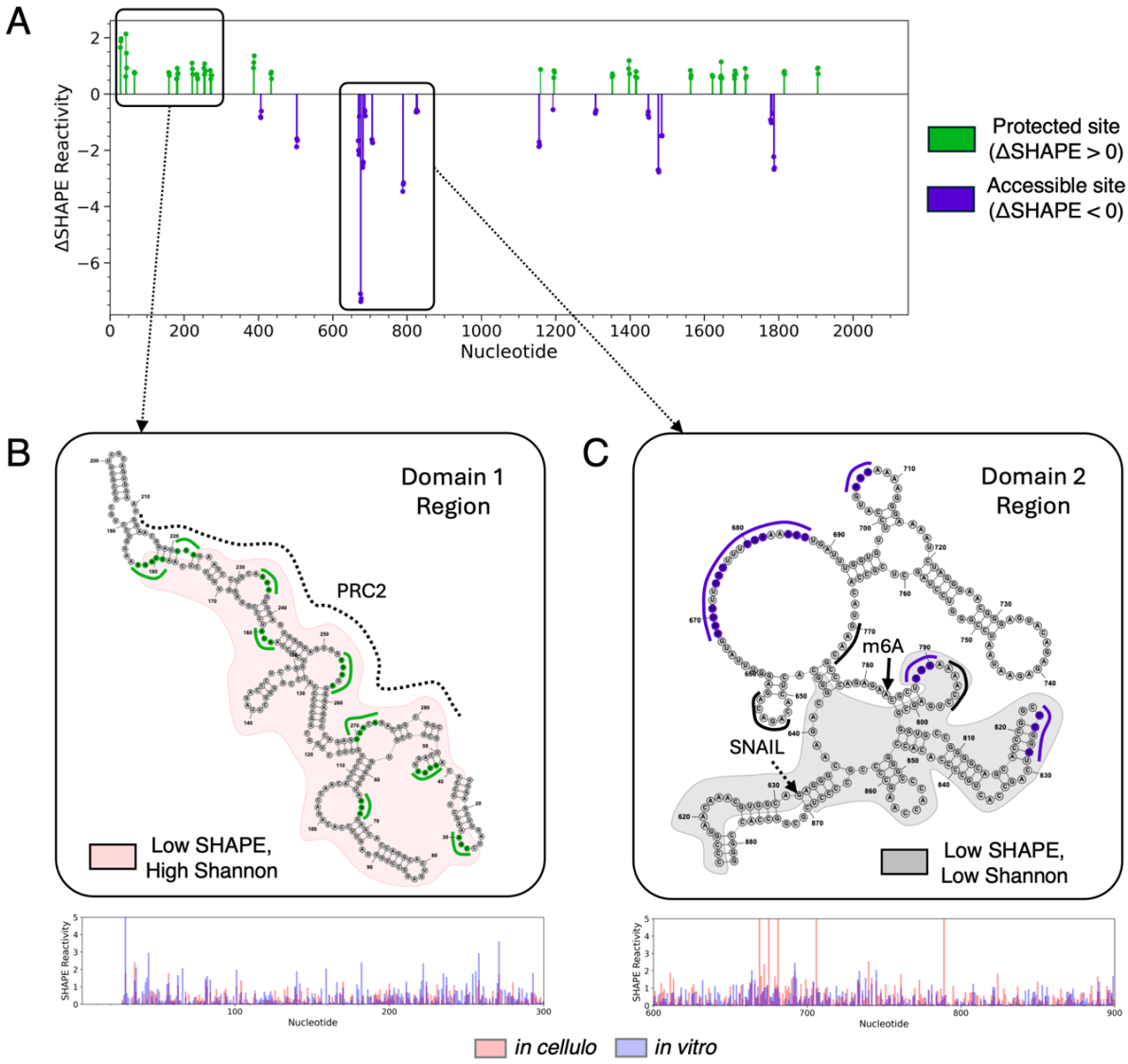



**Figure 2.5 Reactivity differences between *in cellulo* and *in vitro* HOTAIR predict sites of protein binding and structural accessibility.**
(A) Nucleotides with significant reactivity differences calculated from subtracting replicate-averaged *in cellulo* from *in vitro* reactivities using the deltaSHAPE pipeline. Green indicates sites that are more protected *in cellulo* (ΔSHAPE > 0) and purple indicates sites that are more accessible

*in cellulo* (ΔSHAPE < 0). (B) Region of Domain 1 (16-283 nt) with sites of increased protection *in cellulo* (green). This overlaps with a low SHAPE, high Shannon region (light red; 27-182 nt, 224-301 nt) and the minimal PRC2 binding site (black curve; 212-300 nt). Corresponding replicate-averaged reactivity values are shown below the structure. Structure was visualized with StructureEditor. (C) Region of Domain 2 (614-882 nt) with sites of increased accessibility *in cellulo* (purple). This overlaps with a low SHAPE, low Shannon region (gray; 596-645 nt, 788-869 nt), the SNAIL binding site (starting nt indicated with dashed arrow; 632-916 nt), and several m6A sites (black arrow or curve). Corresponding replicate-averaged reactivity values are shown below the structure. Structure was visualized with StructureEditor.

A portion of Domain 2 contains a cluster of seven accessible sites (614-882 nt; **Figure 2.5C**) and is flanked by a highly structured, low Shannon entropy region (**Figure 2.3C**). This locally stable structure surrounded by a more accessible loop region may permit protein interactions or tertiary compaction. This region contains a significant portion of the SNAIL binding domain (632-916 nt) and several experimentally mapped N6-methyladenosine (m6A) modification sites [13, 35]. Finally, Domains 3 and 4 have alternating sites of increased protection or accessibility, and seven protected sites overlap with the LSD1 binding site (1500-2146 nt) [9].

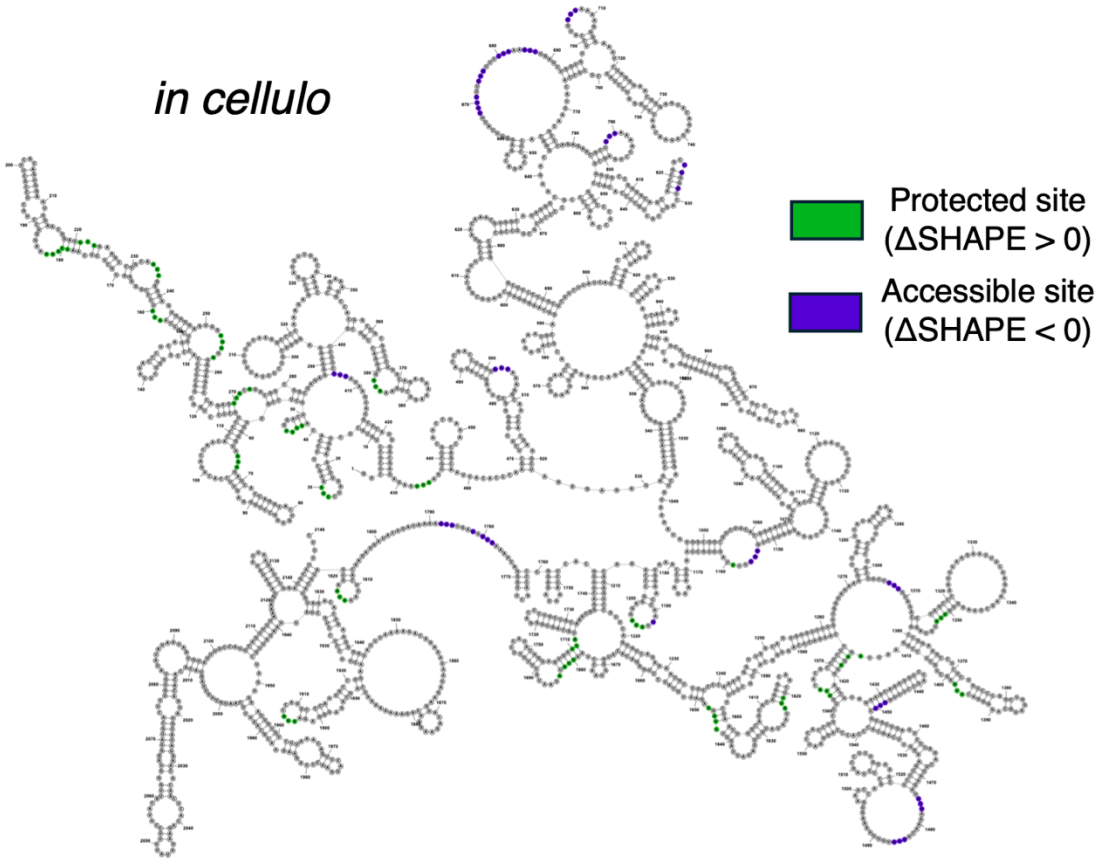



**Figure 2.6 Differential reactivity sites on full-length *in cellulo* HOTAIR.**
Green indicates sites that are more protected *in cellulo* (ΔSHAPE > 0) and purple indicates sites that are more accessible *in cellulo* (ΔSHAPE < 0). Structure was visualized with StructureEditor.

*2.4.5 Conservation and Covariation of HOTAIR in Primates*

Although HOTAIR is present across mammals, its sequence rapidly diverges over evolutionary time [36], so we restricted our analysis to the primate lineage. To better assess conservation but still have sufficient diversity for covariation, we analyzed 190 newly sequenced genomes representing 15 of 16 primate families [37]. Since these genomes are unannotated, we identified candidate loci by searching each genome for homologous sequences to human HOTAIR

and then extracting the best-scoring alignment for each species. Putative HOTAIR homologs were found in all genomes and used to construct a multiple sequence alignment.

Per-nucleotide conservation was assessed across the 191-primate alignment by measuring information content and the frequency of non-gapped positions (**Figure 2.7A**). Regions with both high conservation and high non-gapped frequency indicate segments with high nucleotide identity and alignment coverage. Although conservation was generally maintained among primates, localized insertions and deletions and domain-level conservation differences were observed (**Figure 2.7A**). To quantify domain-level conservation, subalignments of each domain were extracted and the pairwise identity and coverage between each primate species and human were calculated to determine the per-species effective identity. Compared to the per-species effective identities for the full-length alignment, Domain 1 is significantly less conserved while Domains 2-4 show significantly elevated conservation (**Figure 2.7B**). There is also wide sequence variation among individual primates, particularly in Domains 2 and 4.

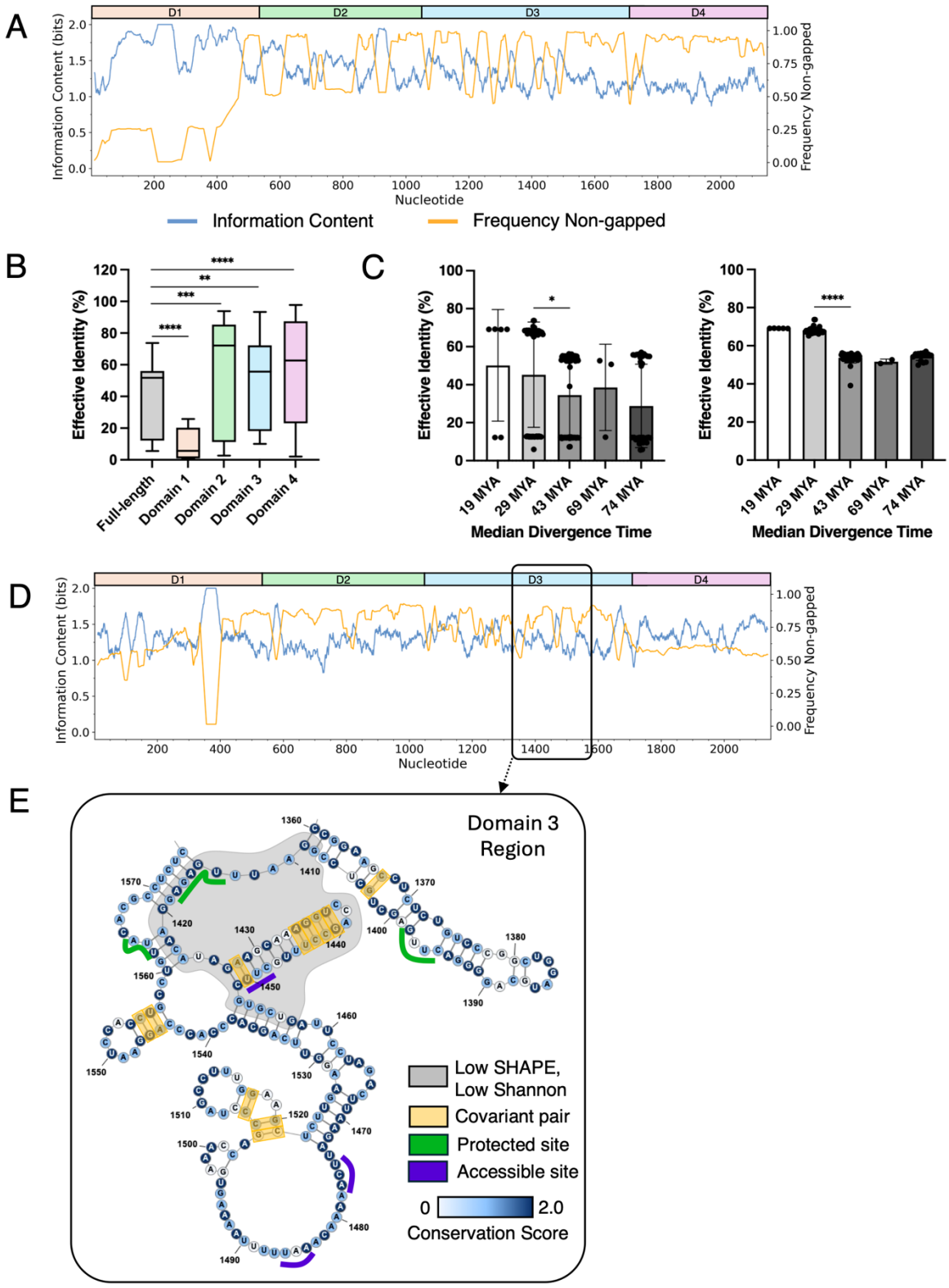



**Figure 2.7 HOTAIR conservation varies by domain and divergence time across primates with evidence for a conserved secondary structure in Domain 3.**

(A) Rolling average of information content and non-gapped frequency across HOTAIR for the unfiltered alignment of 191 primate genomes. Domains 1-4 for the *in cellulo* structure are designated as D1-D4. (B) The pairwise effective identity for each primate compared to human HOTAIR in 190 primate genomes for the full-length and subalignments of Domains 1-4. Visualized with a box-and-whisker plot in GraphPad Prism. (C) The pairwise effective identity for each primate compared to human HOTAIR grouped by median divergence time from human in million years ago (MYA) (left) and only those with >20% effective identity (right). Bar height represents the mean and error bars represent the standard deviation as plotted in GraphPad Prism. (D) Rolling average of information content and non-gapped frequency across HOTAIR for an alignment of 66 primate genomes that was filtered by 95% sequence identity. Domains 1-4 for the *in cellulo* structure are designated as D1-D4. (E) Region of Domain 3 (1360-1574 nt) with covariant base pairs (gold), protected (green) and accessible (purple) sites, and per-nucleotide conservation scores from (D). This overlaps with a low SHAPE, low Shannon region (gray; 1408-1457 nt). Structure was visualized with StructureEditor. Statistical significance calculated with unpaired, two-tailed Student's t-test, **** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, * = $p \leq 0.05$. All alignments include human HOTAIR.

To analyze HOTAIR conservation over evolutionary time, the primate species were grouped based on their median divergence time from human. The pairwise effective identities generally decrease with increasing divergence time, as expected, with a significant decrease occurring between the groups that diverge 29 million years ago (MYA) and 43 MYA (**Figure 2.7C**, left). Generally, the effective identities segment into high (>20% effective identity) and low (<20% effective identity) groups. The high group, which includes 101 genomes with both high identity and coverage, also shows the same conservation trend, although the decrease between the 29 MYA and 43 MYA groups is more significant (**Figure 2.7C**, right). The assembly quality metric N50 shows a wide range of values across the effective identities (**Figure 2.8A**), indicating that the low group includes both sequences with insufficient assembly quality to determine the locus (low N50) and those that may be genuinely divergent (high N50).

To detect covariation, the 191-primate alignment was subsequently filtered to remove highly similar sequences using a 95% sequence identity threshold, which resulted in a 66-primate alignment (including human HOTAIR) (**Figure 2.8B-D, Table 2.3**). Filtering reduced

conservation bias from sampling closely related species (**Figure 2.7D**), whereas more stringent filtering at 85% sequence identity decreased non-gapped frequency and reduced overall alignment quality (**Figure 2.8B**). Per-nucleotide conservation scores from the filtered alignment (**Figure 2.7D**) were visualized on the *in cellulo* structure (**Figure 2.8E**).

One segment with fair conservation and high non-gapped frequency is located within Domain 3 (1360-1574 nt). Although insertions are present in some primate species, conservation is much higher within the segments predicted to be base paired in human HOTAIR (**Figure 2.8F**). Covariation analysis identified eleven significantly covarying base pairs in this segment out of 13 predicted to fall within Domain 3, supporting the presence of a conserved structure in this domain (**Figure 2.7E**). Phylogenetic analysis of this alignment broadly separated the primates into two major groups (**Figure 2.8G**). For the covarying base pairs, base-pair identity was more often maintained within the first primate group, which includes human HOTAIR, and compensatory substitutions were more frequently observed in the second group (**Table 2.4**). Interestingly, this segment overlaps with several protected and accessible sites (**Figure 2.5A**) and a low SHAPE, low Shannon region that was identified in both the *in cellulo* and *in vitro* contexts (**Figure 2.3C-D**). These observations suggest that Domain 3 contains a conserved, stably structured element with functional relevance.

Overall, these results show that HOTAIR maintains a four-domain architecture *in cellulo* with domain-level differences in local structure and accessibility relative to *in vitro*. In addition, we provide evidence for sequence and structural conservation over evolutionary time and within specific domains of HOTAIR.

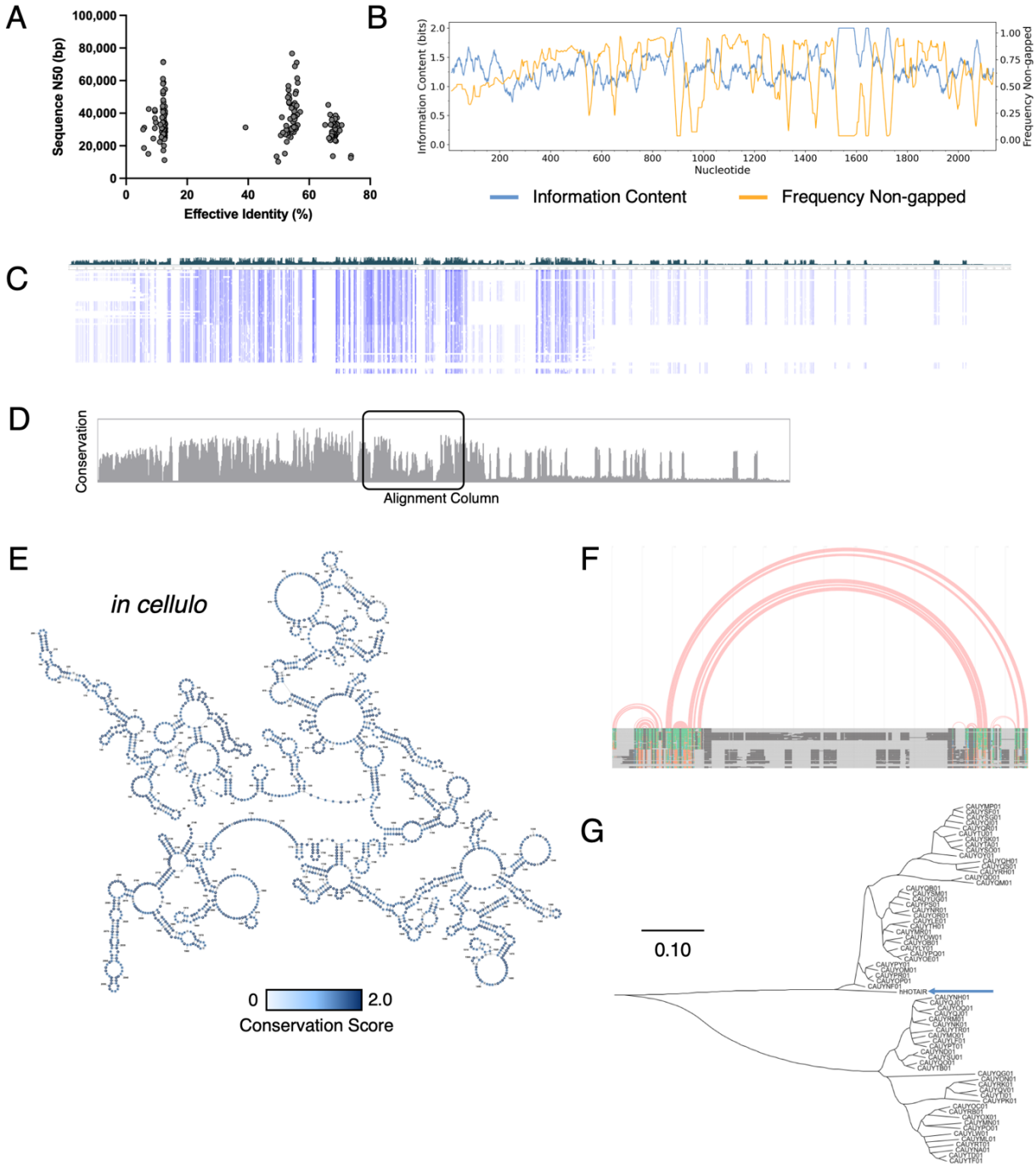



**Figure 2.8 HOTAIR sequence and secondary structure conservation patterns across primates.**

(A) Sequence N50 in base pairs (bp) of each assembly compared to the pairwise effective identity from the unfiltered alignment. Plotted in GraphPad Prism. (B) Rolling average of information content and non-gapped frequency across HOTAIR for an alignment of 27 primate genomes that was filtered by 85% sequence identity. (C) Conservation of HOTAIR across an alignment of 66 primate genomes that was filtered by 95% sequence identity. Each nucleotide is colored by

percentage identity. Visualized with UGENE. (D) Conservation of HOTAIR across an alignment of 66 primate genomes that was filtered by 95% sequence identity. Black box corresponds to the Domain 3 Region. Visualized with UGENE. (E) Per-nucleotide conservation scores from the alignment in (B-D) on full-length *in cellulo* HOTAIR. Structure was visualized with StructureEditor. (F) Portion of alignment corresponding to a region in Domain 3 (1360-1574 nt). Each arc represents a base pair. Nucleotide coloring: conservation (green), covariation (blue), one-sided covariation (light blue), invalid base pair (orange), unpaired (black), and gap (gray). Visualized with Rchie. (G) Phylogenetic tree of primate species in the alignment in (C), where 0.1 = 10% sequence divergence. Tree was built using Maximum Likelihood default parameters in MEGA. Human HOTAIR (hHOTAIR) is indicated with a blue arrow. All alignments include human HOTAIR.

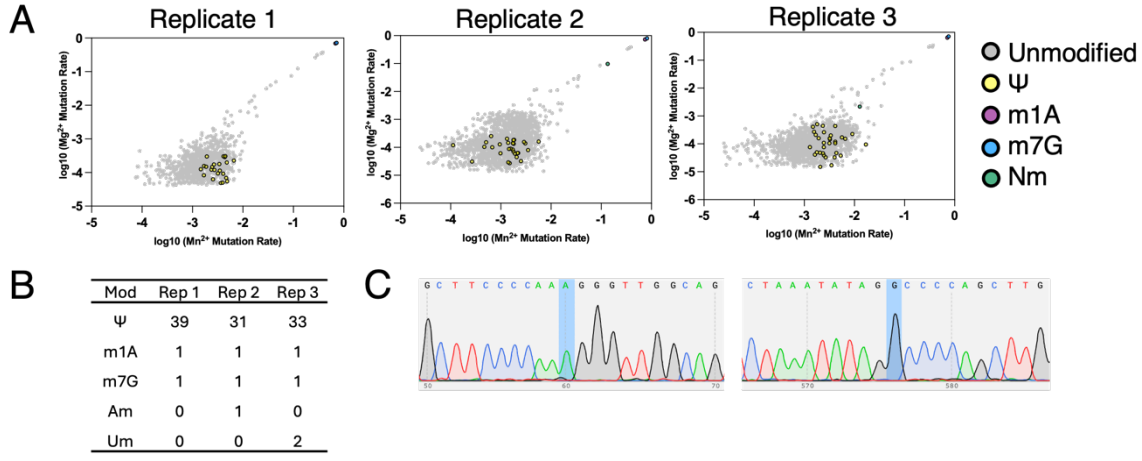
## 2.5 Exploratory Analyses

The results in this section represent exploratory experiments and analyses to gain more insight into the mechanism of HOTAIR in breast cancer cells or identify structural patterns. Although preliminary, these observations can motivate future studies of HOTAIR or other lncRNAs.

### *2.5.1 Discovering Novel RNA Modifications*

To discover potential RNA modifications on HOTAIR in breast cancer cells, the negative control total cellular RNA samples used for SHAPE-MaP were subjected to the MRT-ModSeq pipeline [38]. Briefly, reverse transcription was performed with MarathonRT in the presence of $Mg^{2+}$ or $Mn^{2+}$ cofactors. Samples were amplified and sequenced (See 'RT-PCR, Library Preparation, and Sequencing' in **2.7 Materials and Methods**), and machine learning in combination with mutation signature filtering was used to identify the presence of potential RNA modifications. Across three biological replicates, N1-methyladenosine (m1A, nt 1243) and N7-methylguanosine (m7G, nt 1759) were consistently predicted, and genomic DNA sequencing confirmed that these sites were not single nucleotide polymorphisms (SNPs) (**Figure 2.9A, C**). Other modifications were identified, including ~34 candidate pseudouridine (Ψ) sites per replicate

(**Figure 2.9B**). Pseudouridine predictions tend to have high false positive rates, and only eight Ψ sites were present in at least two biological replicates. The sequences surrounding the candidate Ψ sites were compared to known pseudouridine synthase recognition motifs, but no strong matches were found. These candidate modifications would need to be validated with other methods, such as RNA pull downs coupled with mass spectrometry (for m1A or m7G) or BID-seq (for Ψ) [39]. Although preliminary, modifications such as these may influence HOTAIR's secondary structure or interactions with binding proteins.




| Mod | Rep 1 | Rep 2 | Rep 3 |
|---|---|---|---|
| Ψ | 39 | 31 | 33 |
| m1A | 1 | 1 | 1 |
| m7G | 1 | 1 | 1 |
| Am | 0 | 1 | 0 |
| Um | 0 | 0 | 2 |

**Figure 2.9 Modifications are consistently predicted across biological replicates.**
(A) RNA modifications identified in MDA-MB-231 HOTAIR cells using the MRT-ModSeq pipeline on 3 biological replicates. Plotted in GraphPad Prism. (B) Corresponding table of identified modifications. (C) Sequencing of genomic DNA to confirm m1A1243 and m7G1759 are not SNPs, with a representative replicate shown. Visualized in SnapGene.

### *2.5.2 Breast Cancer Spheroid Culture*

To better model breast cancer tumors, MDA-MB-231 or MDA-MB-231 HOTAIR cells were grown as 3D spheroid cultures [40]. Briefly, cells were seeded onto 96-well ultra-low attachment plates with collagen to enable uniform spheroid formation following an adapted protocol [41]. Z-stack images of individual spheroids were captured, and image thresholding was

applied to quantify spheroid area (Representative spheroids shown in **Figure 2.10**). The initial goal was to develop a functional assay to measure invasion into an extracellular matrix, but no invasion was observed [42]. RT-qPCR measurements of total cellular RNA extracts from spheroid culture indicated sufficiently abundant HOTAIR for SHAPE-MaP, so future studies could determine HOTAIR's secondary structure in 3D culture and compare it to the structure from this dissertation obtained in 2D culture. Although preliminary, morphological differences were observed in the two cell lines: MDA-MB-231 spheroids appeared more compact and uniform, while HOTAIR-overexpressing spheroids were more dispersed and heterogeneous, sometimes forming cavities (See B, C, D in **Figure 2.10**). These observations suggest that 3D culture may be a promising avenue to explore the relationship between HOTAIR structure and function in a more physiologically relevant context.

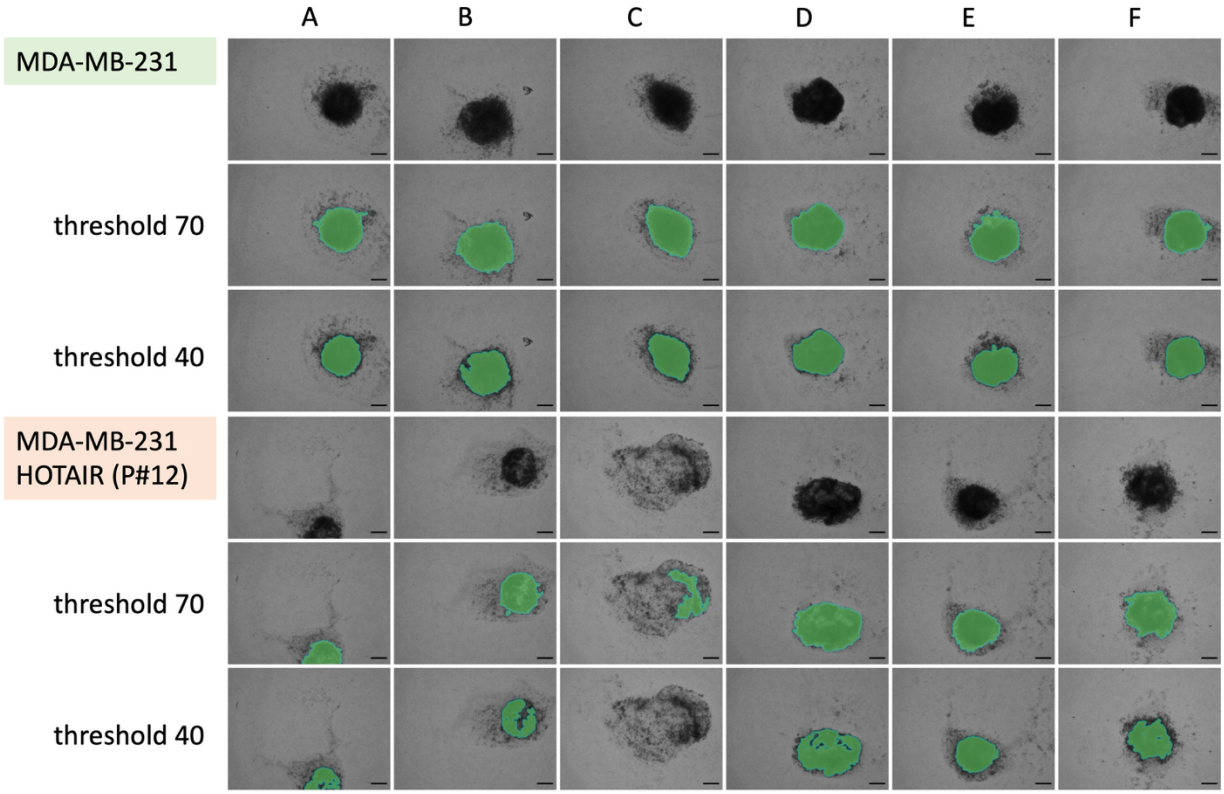



**Figure 2.10 Morphology of spheroids in MDA-MB-231 and MDA-MB-231 HOTAIR Cells.**

MDA-MB-231 (top) and MDA-MB-231 HOTAIR (bottom) cells were seeded at 5,000 (A, B, C) or 10,000 (D, E, F) cells/well. Media was supplemented with 4 μg/mL of collagen. Imaged 4 days post-seeding on a Keyence microscope with thresholds determined using Keyence analysis software. Scale bar = 250 μm. P# = Passage number.

### *2.5.3 RNA Graphs and Nonlinear Metrics*

In addition to standard SHAPE reactivity and Shannon entropy criteria for identifying structured regions, alternative criteria may help identify structural patterns that would otherwise be overlooked. To explore this, HOTAIR's secondary structure was converted into a graph representation in sliding windows across the RNA, with nucleotides represented as nodes and base pairs or backbone connections as edges, inspired by approaches previously applied to proteins and RNA [43]. From these graphs, structure metrics were calculated, including the approximate radius of gyration, which reflects structural compactness, and fractal dimension, which reflects self-similarity or hierarchical organization. The resulting values were converted to Z-scores, and several regions of interest were identified (**Figure 2.11**). Some of the regions identified by both metrics overlapped, and these regions may correspond to especially compact or hierarchical structures. Extensive benchmarking on structured RNAs is necessary to determine how these graph-based metrics compare to linear criteria in identifying functional structural modules.

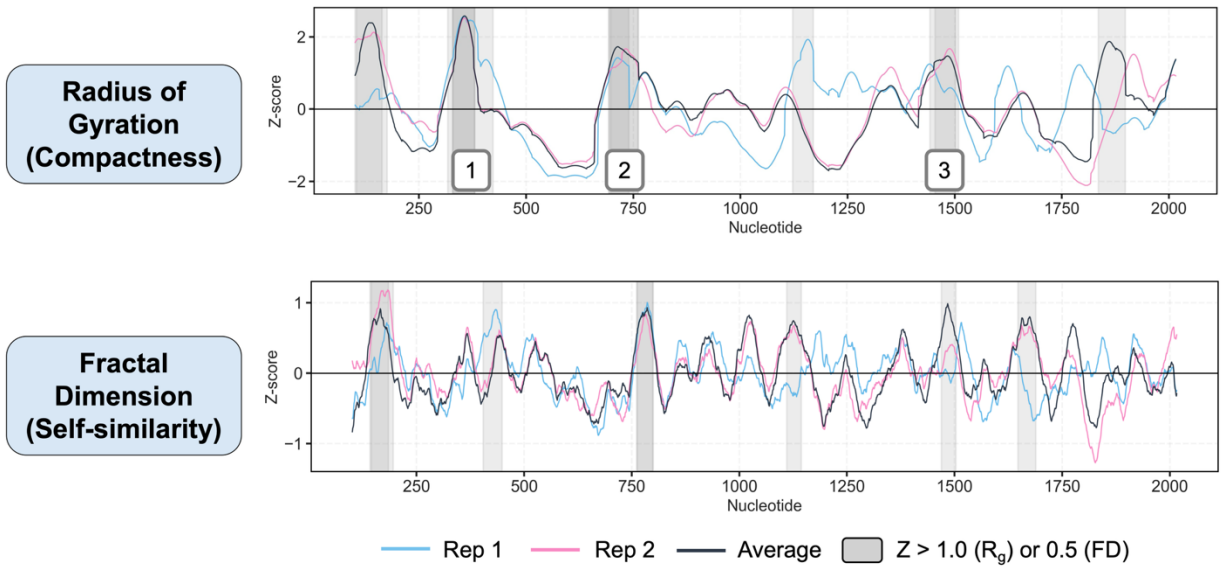



**Figure 2.11 Graph-based metrics to identify structural modules in HOTAIR.**
Radius of gyration (top, $R_g$) and fractal dimension (bottom, FD) were calculated from graphs built in sliding windows of 201 nt with 1 nt steps using the *in cellulo* secondary structure predictions from replicate 1 (blue line), replicate 2 (pink line) or the average (gray line). Regions with elevated values above a Z-score threshold are highlighted in gray. Regions designated as 1-3 had closer agreement between replicates and the average.

## 2.6 Discussion

In this work, we have determined a full-length *in cellulo* secondary structure model of HOTAIR in the metastatic breast cancer cell line MDA-MB-231 using SHAPE-MaP and have generated a corresponding *in vitro* structure for quantitative comparison. The structural models show that HOTAIR adopts a four-domain architecture in both contexts but exhibits domain-specific structural differences in cells. In addition, bioinformatic analyses of newly available primate genomes reveal conservation of the HOTAIR sequence and evidence of structural conservation in specific regions. Together, these results provide a framework for more precise structure-function mechanistic studies of HOTAIR in cellular and disease-relevant contexts.

Comparison of the *in cellulo* and *in vitro* models reveals both shared and context-dependent structural features that may influence HOTAIR function in cells. The consistency of

the four-domain architecture indicates that the global fold is largely encoded by the primary sequence and can form independently of cellular components, such as protein binding partners, chaperones, or molecular crowding. In contrast, structural differences are observed at the individual domain level, suggesting that cellular factors modulate specific structural regions in HOTAIR. For example, there are several low SHAPE reactivity regions that are unique to the *in cellulo* structure. The boundaries of Domains 3 and 4 also differ slightly between contexts, which may be due to the involvement of protein binders, consistent with the fact that these domains overlap with the reported LSD1 binding site (1500-2148 nt) [9]. These observations suggest a general folding principle for long RNAs, where the overall architecture is stably encoded by the sequence, but cellular interactions tune local structural features and domain boundaries. Furthermore, single-stranded 'linker' regions that separate the discrete structures contained within each domain may provide flexibility for tertiary folding.

Well-structured modules are typically defined as regions with both low SHAPE reactivity and low Shannon entropy across a defined nucleotide window [22]. However, this approach only captures the most static structures within an RNA. By conducting a parallel analysis that identifies regions of low SHAPE reactivity corresponding to high Shannon entropy, we can identify highly structured regions that may exhibit conformational heterogeneity, consistent with recent approaches [44]. For example, well-paired RNA structures or switches that sample two or more conformational states could be identified through this approach. Structures of this type are predicted in Domain 1 and 2 *in cellulo*, and regions such as this, which are classified as highly structured but are not necessarily rigid, may represent an important class of RNA motifs for which multiple conformations are essential for function.

Beyond structure determination, directly comparing experimental probing data between *in cellulo* and *in vitro* contexts reveals areas of differential reactivity. Generally, these sites are not evenly interspersed on HOTAIR but tend to form 'hotspots' of either increased protection or accessibility. Notably, there is a large, protected region in Domain 1 that closely matches the reported PRC2 binding site (1-300 nt) [9, 34]. This region overlaps with a low SHAPE, high Shannon region observed *in cellulo*, and we cannot rule out that protein binding may be confounding these measurements. Low SHAPE reactivity may reflect protection due to protein binding rather than a stable structure, while high Shannon entropy may reflect computational uncertainty in the prediction rather than true conformational heterogeneity.

Conversely, the SNAIL binding domain overlaps with clusters of increased accessibility in Domain 2, suggesting that these types of sites may also facilitate protein binding. The low SHAPE, low Shannon region here may help fold the structure to create a modular platform for protein binding. Interestingly, the mapped m6A sites in this region all correspond to single-stranded regions, making them accessible to modification-interacting enzymes. This region also exhibits distinct base pairing from the *in silico* and *in vitro* models (**Figure 2.4G**), making it an interesting structural target unique to the cellular context. Overall, HOTAIR's structure is consistent with protein interactions that promote epigenetic silencing and cancer progression. We hope that these findings provide a structural foundation to guide future studies of the HOTAIR-protein interactome in cancer.

Because lncRNAs often diverge rapidly in sequence, restricting conservation analyses to primates allows for better assessment of conservation. Homologous HOTAIR sequences were identified across all the genomes, although incomplete assemblies and unannotated genomes limit definitive identification of the HOTAIR loci. Synteny analysis with neighboring genes on the

*HOXC* locus on several high-quality primate genomes is necessary to confirm the correct locus identification. Despite these limitations, our alignments reveal substantial sequence conservation across portions of the HOTAIR transcript, particularly in Domains 2 and 4, which overlap with the SNAIL and LSD1 binding domains, respectively. Domain 1, which overlaps with six exons in HOTAIR, shows lower conservation and higher gapped frequency relative to the other domains, which may reflect challenges in extracting the locus across exon boundaries, variation in exon composition in some species, or incomplete assemblies. Since Domain 1 includes the PRC2 binding domain, this functionality may be lost in some primates.

Within the primate lineage, the overall HOTAIR sequence rapidly diverges across evolutionary time, particularly from 29 MYA to 43 MYA. The 19 MYA and 29 MYA divergence groups represent Hylobatidae (lesser apes) and Cercopithecidae (Old World monkeys), respectively, while the 43 MYA, 69 MYA, and 74 MYA groups represent Platyrrhini (New World monkeys), Tarsiidae (tarsiers), and Strepsirrhini (lemurs, lorises, and galagos). Changes in conservation across these primate clades may partly reflect the influence of geographic dispersal as a driver of HOTAIR sequence divergence.

Taking advantage of the 190 primate genomes allowed inclusion of novel species and provided sufficient phylogenetic diversity for covariation analysis. We show strong evidence for covariant base pairs in Domain 3, supporting the functional relevance of a discrete structural element. This element overlaps with a low SHAPE reactivity and low Shannon entropy region and is close to several protected and accessible sites, all of which are indicators of potential functional importance. Several covariant base pairs were predicted in the other domains as well using the filtered alignment (3 pairs in Domain 1, 2 pairs in Domain 2, and 7 pairs in Domain 4).

By limiting our analysis to primate genomes, we show that HOTAIR is far more conserved than generally appreciated, and selective pressure may act on specific structural elements.

A limitation of this study is that HOTAIR overexpression was required to obtain sufficient transcript abundance for probing. Although the overexpression levels are comparable to some patient tumors, the structure presented here may not fully recapitulate native folding in the tumor environment. Future studies that probe endogenous HOTAIR, such as in patient-derived xenograft cell lines [45] or primary breast tumor samples [15], would help address this limitation. Additionally, chemical probing reports indirectly on RNA structure and captures an ensemble-averaged snapshot, rather than distinct conformational states. RNAs, like proteins, likely adopt multiple conformations, some of which may guide distinct cellular functions. Future studies in which libraries are constructed from specific subcellular compartments can address this issue. Finally, identification and alignment of lncRNAs across species remains challenging due to frequent insertions and deletions, exon variation, and incomplete genome assemblies, which limits our ability to detect covariation. Addressing these challenges will allow future studies to make critical links between structure and function in HOTAIR and other lncRNA targets.

The *in cellulo* secondary structure model of HOTAIR presented here provides a roadmap for future mechanistic studies. Regions that are highly structured and show increased protection, conservation, or covariation represent promising candidates for investigating protein binding and the role of HOTAIR's structure in gene regulation and cancer. More broadly, determining structures of lncRNAs in relevant cellular contexts is a key step for connecting structure to function.

## 2.7 Materials and Methods

### Tissue Culture

MCF7 (ATCC HTB-22) and MDA-MB-231 (ATCC HTB-26) cells were obtained from the American Type Culture Collection. MDA-MB-231 pBABE-puro HOTAIR overexpression cells (MDA-MB-231 HOTAIR) had been previously generated from parental MDA-MB-231 cells using retroviral transduction as described [12, 13]. Cells were grown in Dulbecco's Modified Eagle Medium with high glucose and L-glutamine without sodium pyruvate (GenClone, Gibco) with 10% fetal bovine serum (Avantor) and 1% penicillin-streptomycin (Gibco) at 37ºC and 5% $CO_2$. Media for MDA-MB-231 HOTAIR cells was supplemented with 1 µg/mL puromycin (Gibco).

### *In vitro* Transcription and Purification for qPCR

*In vitro Transcription*. HOTAIR was transcribed from a pCARs08 plasmid containing the HOTAIR sequence immediately downstream of a T7 promoter. The plasmid had been previously generated as described [24]. Plasmid was linearized by incubating with SalI restriction enzyme and CutSmart Buffer at 37ºC for 16 hrs, followed by enzyme inactivation at 95ºC for 5-10 min. Plasmid was purified with the QIAquick Gel Extraction Kit (Qiagen) and added to a 1 mL transcription reaction with transcription buffer (15 mM MgCl2, 40 mM Tris-Cl pH 8, 2 mM spermidine, 10 mM NaCl, 0.01% Triton 100), 10 mM DTT, 4 mM of each NTP at pH 7, RNase OUT (Invitrogen), and 0.08 mg/mL T7 enzyme (purified *in house*). Reaction was incubated at 37ºC for 2 hrs, followed by plasmid digestion with 4 µL TURBO DNase (Invitrogen) at 37ºC for 30 min and addition of 10 µl of 0.5 M $Na_2EDTA$ pH 8.5. Transcribed RNA was concentrated to

500 µL and exchanged into 150 mM KCl and 50 mM MOPS pH7 using a 100 kDa Amicon Ultra 0.5 mL filtration column (Amicon).

*Gel Purification*. Urea loading dye (11 M urea, 2% xylene cyanol/bromophenol blue, 50% sucrose, 40 mM Tris-HCl pH 7.5, 0.8 mM $Na_2EDTA$ pH 8.5 pH 8) was heated at 65ºC for 10 min and added to the buffer-exchanged RNA in a 1:1 ratio. An 8 M Urea-5% Acrylamide-Tris-Borate-EDTA (TBE) denaturing gel was cast onto custom-made siliconized gel plates from the Yale University Scientific Glassblowing Laboratory and polymerized with ammonium persulfate and TEMED. The gel was run at 20 W for 30 min prior to loading the RNA sample, which was run at 20-25 W for 3 hours. Gel plates were opened and plastic wrap was laid on top of the gel to flip onto a gel tray. RNA was then visualized under UV light, and the band was extracted with a sterile blade. The RNA was eluted with the 'crush and soak' method. Briefly, the gel slice was forced through a 5 ml syringe, 10-12 mL of elution buffer (300 mM NaCl, 10 mM MOPS pH 6, 1 mM $Na_2EDTA$ pH 8.5) was added, and RNA was eluted at 4ºC overnight with shaking. Gel pieces were removed with a filter. 3 M sodium-acetate pH 5.2 was added to precipitate the RNA at -80ºC for 2 hrs. To obtain a pellet, the sample was spun at 15,000g for 30 min, washed with 70% EtOH, and spun at 15,000g for another 30 min. The pellet was allowed to dry and redissolved in 1x ME storage buffer (10 mM MOPS pH 6.5, 1 mM $Na_2EDTA$ pH 8.5). Concentration was determined by Nanodrop at A260.

**Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR)**

*Reverse Transcription (RT)*. To generate a standard curve of *in vitro* transcribed HOTAIR, a 10-fold serial dilution of $10^{11}$ copies/µL to $10^3$ copies/µL was made using the molecular weight

calculated from AAT Bioquest [46]. For total cellular RNA samples, cells were dislodged with cell scrapers, centrifuged, and resuspended in Dulbecco's PBS (GenClone, Gibco). RNA was extracted using 3 parts TRIzol (Invitrogen) to 1 part cell suspension followed by adding chloroform:isoamyl alcohol 24:1 (Sigma) equal to 1/5 TRIzol volume. After centrifuging at 3,000g for 15 min at 4ºC, the aqueous RNA phase was removed and precipitated in 75% final EtOH at -80ºC for several hours. RNA was pelleted at 15,000g for 30 min at 4ºC, air dried, and resuspended in 1x ME buffer. Samples were treated with DNase I (Qiagen) prior to clean-up with the RNeasy kit (Qiagen). For annealing, 1 µg of each RNA sample was combined with 10 mM dNTP Mix (NEB), 2 mM RT primer, and RNase-free water and incubated at 65ºC for 5 min in a PCR thermocycler, then placed on ice for 1 min. To the annealing mix, 5x First Strand Synthesis buffer (Invitrogen), 10 mM DTT, 1 µL SUPERase•In RNase inhibitor (Invitrogen), and 1 µL Superscript (SS) II or III (Invitrogen) were added for a final RT reaction volume of 20 µL. Reverse transcription proceeded at 55ºC for 50 min for SSIII or 42ºC for 50 min for SSII followed by enzyme inactivation at 85ºC for 5 min. Template RNA was degraded via alkaline hydrolysis in which 1 µL 3 M KOH was added, the solution was incubated at 95ºC for 5 min and 4ºC for 2 min, and then 1 µL 3 M HCl was added. The cDNA was cleaned with 1.8x AMPure XP beads (Beckman) according to the manufacturer's protocol.

*Quantitative Polymerase Chain Reaction* (*qPCR)*. For each qPCR reaction, 5 µL cleaned cDNA, 2x Light Cycler 480 SYBR Green I Master Mix (Roche), 10 µM forward primer, 10 µM reverse primer, and RNase-free water was combined for a 20 µL reaction. The qPCR plate was covered with a plastic sheet and centrifuged at 2,000g for 2 min. Plates were immediately placed in the

qPCR thermocycler and subjected to pre-incubation, amplification (40 cycles), melting curve, and cooling steps.

*qPCR Analysis*. Technical duplicates and/or biological replicates were averaged. The standard curve was plotted and fit to a linear regression line. The slope and y-intercept were used to determine the copies/µl of HOTAIR in cellular samples and converted to copies/cell based on the volume of RNA per dish and the cell count determined using a hemocytometer and Trypan Blue (Sigma). All primers for qPCR are listed in **Table 2.2**.

### *In cellulo* SHAPE Probing

*Harvesting Cells*. MDA-MB-231 HOTAIR cells were grown to 80-90% confluency on two 150 mm dishes per biological replicate. The media was aspirated, the cells were washed with phosphate-buffered saline (PBS), and 10 mL PBS was added to each dish. Cells were dislodged with a cell scraper, pooled, and centrifuged at 1,000g for 5 min at 4°C. Cell pellets were resuspended in 1.5 mL PBS, then evenly split into treated and control fractions.

*In cellulo SHAPE Probing*. Solution of 1 M 2A3 was prepared immediately prior to probing by dissolving 100 mg of 2A3 into 530 µL DMSO (Sigma) and protected from light. For the 2A3-modified fraction, 100 µL of 1 M 2A3 was added per 0.5 mL cell suspension for a final concentration of 100 mM and mixed well. For the DMSO-unmodified control fraction, 100 µL of 100% DMSO was added per 0.5 mL cell suspension and mixed well. Both fractions were incubated at 37ºC for 15 minutes protected from light.

*Total Cellular RNA Extraction.* To quench the reaction, 3 parts TRIzol (Invitrogen) were added to 1 part cell suspension, mixed well, and incubated for 5 min at room temperature. Next, chloroform:isoamyl alcohol 24:1 (Sigma) equal to 1/5 TRIzol volume was added, mixed well, and incubated for 5 min at room temperature. Samples were centrifuged at 4,700 rpm for 15 min at 4°C, and the aqueous RNA phase was precipitated with 1 µL glycogen per 1 mL aqueous RNA and 75% final EtOH at -20°C for 2-3 days.

*EtOH Precipitation and RNA Cleanup*. RNA was pelleted by spinning at 15,000g for 30 min at 4°C. After air drying, each RNA pellet was resuspended in RNase-free water, and in-solution DNase digestion was performed with TURBO DNase buffer and 5 µL TURBO DNase (Invitrogen) at 37ºC for 30 min. DNase-treated RNA was purified using the RNeasy kit (Qiagen) and eluted in 1x ME buffer (10 mM MOPS pH 6.5, 1 mM $Na_2EDTA$ pH 8.5) supplemented with SUPERase•In RNase inhibitor (Invitrogen). Concentration was determined by Nanodrop at A260.

### *In vitro* SHAPE Probing

*In vitro Transcription*. HOTAIR pCARs08 plasmid was linearized with SalI-HF restriction enzyme (NEB) and CutSmart Buffer at 37ºC for 1 hr with no heat inactivation. Plasmid was precipitated with 1/10 volume of 3 M NaOAc pH 5.5 and 75% EtOH at -20ºC for 1 hr. Plasmid was pelleted by spinning at 21,000g for 30 min at 4ºC, washed with 70% EtOH, and spun at 21,000g for 5 min. Pellet was allowed to air dry before resuspending in RNase-free water. *In vitro* transcription was performed with transcription buffer (15 mM MgCl2, 40 mM Tris-Cl pH 8, 2 mM spermidine, 10 mM NaCl, 0.01% Triton 100), 10 mM DTT, 2.5 mM of each NTP at pH 7, SUPERase•In RNase inhibitor (Invitrogen), thermostable inorganic pyrophosphatase (NEB), 10

mM $MgCl_2$ (Invitrogen) and 0.2 mg/mL T7 enzyme (purified *in house*) for a 1 mL reaction. Reaction was incubated at 30ºC for 1.5 hrs, followed by plasmid digestion with 20 µL TURBO DNase (Invitrogen) at 37ºC for 10 min and addition of 20 µl of 0.5 M $Na_2EDTA$ pH 8.5. Transcribed RNA was concentrated and exchanged into filtration buffer (50 mM K-HEPES pH 7.2, 150 mM KCl, 0.1 mM $Na_2EDTA$ pH 8.5) with 100 kDa Amicon Ultra 0.5 ml filtration columns (Amicon).

*Semi-native Purification and Folding*. RNA was semi-natively purified with a fast performance liquid chromatography AKTA using a Sephacryl S-400 column (Cytiva) equilibrated with filtration buffer. RNA elution into 0.5 mL fractions was monitored using the absorbance at A260. The peak fraction was collected, and concentration was determined with Nanodrop at A260. For folding, $MgCl_2$ was added to a final concentration of 25 mM and incubated at 37°C for 30 min.

*In vitro SHAPE Probing*. Immediately after folding, the sample was split into 5 µg fractions for each probing condition and replicate. To modified fractions, 2A3 was added for a final concentration of 100 mM. An equivalent volume of DMSO (Sigma) was added to unmodified control fractions. Both fractions were mixed well and incubated at 37°C for 10 min protected from light. To quench the reaction, 1/10 volume of 3 M NaOAc and 75% EtOH were added to each sample and precipitated at -20ºC overnight. To pellet RNA, samples were spun at 21,100g for 30 min at 4ºC, washed with 70% EtOH, and spun at 21,100g for 5 min at 4ºC. Residual EtOH was removed, and the pellet was air dried and resuspended in RNase-free water. Concentration was determined by Nanodrop at A260.

**RT-PCR, Library Preparation, and Sequencing**

*RT-PCR*. All RNA samples from *in cellulo* and *in vitro* SHAPE probing were reverse transcribed with MarathonRT [47]. Annealing mixes were prepared with 1 µg RNA and 1 µM HOTAIR-specific primer and annealed at 68°C for 5 min. To each reaction, 2.5x MarathonRT buffer (125 mM Tris-HCl pH 7.5, 500 mM KCl, 12.5 mM DTT, 1.25 mM dNTPs (NEB, Thermo), 2.5 mM $MnCl_2$ (Sigma)), MarathonRT (purified *in house*), and glycerol were added and incubated at 42°C for 3 hours, followed by enzyme inactivation at 70°C for 15 min. Single-stranded cDNA was cleaned with either 1.8x RNAClean XP or 1.8x AMPure XP beads (Beckman). A single amplicon from 7-2122 nt (*in cellulo* samples) or two overlapping amplicons from 7-1213 nt and 1010-2122 nt (*in vitro* samples) was amplified with PCR using Q5 High-Fidelity 2X MasterMix (NEB), 10 µM forward primer, 10 µM reverse primer, and 5 µL cleaned cDNA with touchdown PCR conditions. Annealing temperatures were estimated using SnapGene or the NEB Tm Calculator. PCR products were cleaned with 1.8X AMPure XP beads (Beckman). All primers for SHAPE-MaP are listed in **Table 2.2**.

*Library Preparation and Sequencing*. DNA concentration was measured with the Qubit dsDNA HS Assay Kit (Invitrogen) and diluted to 0.2 ng/µL. Tagmentation and Nextera PCR were carried out according to the Nextera XT DNA Library Prep documentation (Illumina), except at 1/5 the volume. Libraries were purified with 1.8X AMPure XP beads (Beckman) and analyzed on a Bioanalyzer with the High Sensitivity DNA Kit or TapeStation with the HS D5000 ScreenTape Assay (Agilent). Final libraries were adjusted to 4 nM, pooled based on the desired read depth, and adjusted to the final loading concentration required for each kit. Sequencing was performed

with the NextSeq 1000/2000 P1 Reagents Kit or the NextSeq 1000/2000 P1 XLEAP-SBS Reagent Kit (Illumina) on a NextSeq 1000/2000.

**Sequencing Alignment and Structure Prediction**

Sequencing files were retrieved using the BaseSpace Sequence Hub Downloader and transferred to an HPC cluster. Paired-end reads were aligned to the full-length HOTAIR sequence, excluding nucleotides up to and including the primer binding sites, using ShapeMapper2 under default parameters [32]. After verifying that samples met sufficient quality control thresholds, the ShapeMapper2 output files (.map) were used for secondary structure prediction with SuperFold under default parameters [23]. The *in silico* structure prediction was carried out identically, except the SHAPE reactivity values in the .map file were set to -999 to indicate no experimental data.

**Identification of Structured Regions**

The rolling median of normalized SHAPE reactivity (from ShapeMapper2) and Shannon entropy (from SuperFold) values was calculated in sliding windows of 51 nt around the center nucleotide with step size of 1 nt and truncated edge windows. The global median was subtracted from each rolled value, and regions were identified where the resulting SHAPE reactivity and Shannon entropy were both $< 0$ for a minimum length of 40 nt (low SHAPE, low Shannon), indicating highly structured regions that adopt a dominant conformation. Alternatively, regions where SHAPE reactivity was $< 0$ and Shannon entropy was $> 0$ for a minimum length of 40 nt were also identified (low SHAPE, high Shannon), suggesting these regions are highly structured yet conformationally heterogeneous.

**Jaccard Similarity Calculation**

The Jaccard similarity was calculated between each pair of structures in sliding windows of 301 nt around the center nucleotide with step size of 1 nt and truncated edge windows. The Jaccard similarity is defined as the number of shared base pairs (intersection) divided by the total number of unique base pairs (union) in each window.

**ΔSHAPE Analysis**

Replicate-averaged *in cellulo* SHAPE reactivities were subtracted from *in vitro* SHAPE reactivities using the deltaSHAPE pipeline under default parameters and statistical thresholds [33] after converting the deltaSHAPE.py script to Python 3 compatibility.

**Gene Discovery and Multiple Sequence Alignment**

Primate genomes were downloaded from the European Nucleotide Archive (Accession PRJEB67744), deposited by Kuderna et al [37]. For each genome, the tool exonerate (version 2.4.0) with the model est2genome ('model') [48] was used to find potential matches to the human HOTAIR sequence. The match with the best alignment score was extracted using a custom Biopython script. All RNA sequences were aligned using MAFFT (with the 'auto' flag, version 7.526) [49]. The resulting alignment was filtered by 85% or 95% sequence identity using CD-HIT-EST (version 4.8.1) with a word size of 8 ('n') [50, 51] and realigned with MAFFT. The esl-alistat package in HMMR (version 3.4, http://hmmer.org/) was used to calculate per-column information content in bits and non-gapped frequency scores in the input alignment. Per-nucleotide scores in the HOTAIR sequence were extracted, and the rolling average was plotted in sliding windows of 21 nt around the center nucleotide with step size of 1 nt. Alignments were

visualized in UGENE (version 52.0). The maximum likelihood tree was calculated on the 95% filtered alignment in MEGA (version 12.1.1) using default parameters.

**Evolutionary Conservation Analysis**

For conservation analyses, the unfiltered 191-primate alignment (including human HOTAIR) was used. The pairwise identity and coverage of each sequence compared to the human HOTAIR sequence was calculated using the NCBI MSA Viewer (version 1.26.0). We define 'effective identity' as the percent identity multiplied by the coverage. For domain conservation analysis, the effective identity was recalculated for subalignments taken from the full-length alignment corresponding to Domains 1-4. For evolutionary divergence analysis, the median pairwise divergence time between each primate species and human was retrieved from TimeTree [52]. Divergence groups that had at least 3 species were analyzed (thus, *Gorilla beringei* was excluded as the only member of its divergence group). For quality control, the sequence N50 in base pairs of each assembled genome was downloaded [53] and plotted against the effective identity.

**Covariation Analysis**

Covariant base pairs were identified using R-scape (version 0.2.1) installed from Bioconda (https://anaconda.org/bioconda/rscape/files/manage?page=2) and using the RAFSp statistical test [54, 55]. Due to the longer length of HOTAIR, covariation was assessed in sliding windows of 500 nt ('window') with step size 100 nt ('slide'). Input alignments (generated as described above) were converted to Stockholm format using the esl-reformat package in HMMR (version 3.4, http://hmmer.org/), and the replicate-averaged *in cellulo* structure was used. No gaps were

removed from the alignment, so nucleotide positions output from R-scape were converted to the position in the non-gapped human HOTAIR sequence.

**Plotting and Statistical Analysis**

Unless otherwise noted, all plots were generated using Matplotlib or Seaborn in Jupyter Notebook or JupyterLab. For box-and-whisker plots, boxes indicate the interquartile range (IQR), the center line is the median, and the whiskers extend to the farthest data point within 1.5xIQR. Statistical significance was calculated with a two-tailed Mann-Whitney U test or an unpaired, two-tailed Student's t-test, where **** = $p \leq 0.0001$, *** = $p \leq 0.001$, ** = $p \leq 0.01$, and * = $p \leq 0.05$.

**Human HOTAIR Sequence**

The human HOTAIR sequence used in this study comprises nucleotides 1-2148 (excluding the polyA tail) from GenBank Accession DQ926657.1, deposited by Rinn et al. [8], and is listed in 2.8 SUPPLEMENTARY DATA. Exon boundaries were determined from NCBI Reference NR_003716.4; note that this reference is missing the first 5 nt and has a slightly longer final exon compared to the original HOTAIR sequence. Exon 2, 1-60 nt; Exon 3, 61-186 nt; Exon 4, 187-288 nt; Exon 5, 289-412 nt; Exon 6, 413-465 nt; Exon 7, 466-2148 nt. Exon 1 is present in some annotated isoforms and is separated by a relatively long intron.

## 2.8 SUPPLEMENTARY DATA

**Human HOTAIR Sequence**

GACUCGCCUGUGCUCUGGAGCUUGAUCCGAAAGCUUCCACAGUGAGGACUGCUCC
GUGGGGGUAAGAGAGCACCAGGCACUGAGGCCUGGGAGUUCCACAGACCAACACC
CCUGCUCCUGGCGGCUCCCACCCGGGGCUUAGACCCUCAGGUCCCUAAUAUCCCG
GAGGUGCUCUCAAUCAGAAAGGUCCUGCUCCGCUUCGCAGUGGAAUGGAACGGA

UUUAGAAGCCUGCAGUAGGGGAGUGGGGAGUGGAGAGAGGGAGCCCAGAGUUAC
AGACGGCGGCGAGAGGAAGGAGGGGCGUCUUUAUUUUUUUAAGGCCCCAAAGAG
UCUGAUGUUUACAAGACCAGAAAUGCCACGGCCGCGUCCUGGCAGAGAAAAGGCU
GAAAUGGAGGACCGGCGCCUUCCUUAUAAGUAUGCACAUUGGCGAGAGAAUUAA
GUGCUGCAACCUAAACCAGCAAUUACACCCAAGCUCGUUGGGGCCUAAGCCAGUA
CCGACCUGGUAGAAAAAGCAACCACGAAGCUAGAGAGAGAGCCAGAGGAGGGAA
GAGAGCGCCAGACGAAGGUGAAAGCGAACCACGCAGAGAAAUGCAGGCAAGGGA
GCAAGGCGGCAGUUCCCGGAACAAACGUGGCAGAGGGCAAGACGGGCACUCACAG
ACAGAGGUUUAUGUAUUUUUAUUUUUUAAAAUCUGAUUUGGUGUUCCAUGAGGA
AAAGGGAAAAUCUAGGGAACGGGAGUACAGAGAGAAUAAUCCGGGUCCUAGCUC
GCCACAUGAACGCCCAGAGAACGCUGGAAAAACCUGAGCGGGUGCCGGGGCAGCA
CCCGGCUCGGGUCAGCCACUGCCCCACACCGGGCCCACCAAGCCCCGCCCCUCGCG
GCCACCGGGGCUUCCUUGCUCUUCUUAUCAUCUCCAUCUUUAUGAUGAGGCUUGU
UAACAAGACCAGAGAGCUGGCCAAGCACCUCUAUCUCAGCCGCGCCCGCUCAGCC
GAGCAGCGGUCGGUGGGGGGACUGGGAGGCGCUAAUUAAUUGAUUCCUUUGGAC
UGUAAAAUAUGGCGGCGUCUACACGGAACCCAUGGACUCAUAAACAAUAUAUCU
GUUGGGCGUGAGUGCACUGUCUCUCAAAUAAUUUUUCCAUAGGCAAAUGUCAGA
GGGUUCUGGAUUUUUAGUUGCUAAGGAAAGAUCCAAAUGGGACCAAUUUUAGGA
GGCCCAAACAGAGUCCGUUCAGUGUCAGAAAAUGCUUCCCCAAAGGGUUGGCAGU
GUGUUUUGUUGGAAAAAAGCUUGGGUUAUAGGAAAGCCUUUCCCUGCUACUUGU
GUAGACCCAGCCCAAUUUAAGAAUUACAAGGAAGCGAAGGGGUUGUGUAGGCCG
GAAGCCUCUCUGUCCCGGCUGGAUGCAGGGGACUUGAGCUGCUCCGGAAUUUGAG
AGGAACAUAGAAGCAAAGGUCCAGCCUUUGCUUCGUGCUGAUUCCUAGACUUAA
GAUUCAAAAACAAAUUUUUAAAAGUGAAACCAGCCCUAGCCUUUGGAAGCUCUU
GAAGGUUCAGCACCCACCCAGGAAUCCACCUGCCUGUUACACGCCUCUCCAAGAC
ACAGUGGCACCGCUUUUCUAACUGGCAGCACAGAGCAACUCUAUAAUAUGCUUAU
AUUAGGUCUAGAAGAAUGCAUCUUGAGACACAUGGGUAACCUAAUUAUAUAAUG
CUUGUUCCAUACAGGAGUGAUUAUGCAGUGGGACCCUGCUGCAAACGGGACUUU
GCACUCUAAAUAUAGGCCCCAGCUUGGGACAAAAGUUGCAGUAGAAAAAUAGAC
AUAGGAGAACACUUAAAUAAGUGAUGCAUGUAGACACAGAAGGGGUAUUUAAAA
GACAGAAAUAAUAGAAGUACAGAAGAACAGAAAAAAAAUCAGCAGAUGGAGAUU
ACCAUUCCCAAUGCCUGAACUUCCUCCUGCUAUUAAGAUUGCUAGAGAAUUGUGU
CUUAAACAGUUCAUGAACCCAGAAGAACGCAAUUUCAAUGUAUUUAGUACACAC
ACAGUAUGUAUAUAAACACAACUCACAGAAUAUAUUUUCCAUACAUUGGGUAGG
UAUGCACUUUGUGUAUAUAUAAUAAUGUAUUUUCCAUGCAGUUUUAAAAUGUAG
AUAUAUUAAUAUCUGGAUGCAUUUUC

**Table 2.2 List of primers for HOTAIR.**

| Primer | Sequence (5’ to 3’) |
|---|---|
| RT_486 for qPCR | TTCTACCAGGTCGGTACTGG |
| qPCR_14 Forward [13] | TCTGGAGCTTGATCCGAAAG |

| | |
|---|---|
| qPCR_92 Reverse [13] | GGTGTTGGTCTGTGGAACT |
| RT_1196 for SHAPE-MaP [24] | CGGACTCTGTTTGGGCCTCCT |
| RT_2106 for SHAPE-MaP | ATCTACATTTTAAAACTGC |
| PCR_7 Forward for SHAPE-MaP | CCTGTGCTCTGGAGCTTGAT |
| PCR_1010 Forward for SHAPE-MaP | AGGCGCTAATTAATTGATTCC |
| PCR_1196 Reverse for SHAPE-MaP [24] | ACTCTGTTTGGGCCTCCT |
| PCR_2102 Reverse for SHAPE-MaP | CTACATTTTAAAACTGCATGG |

RT, reverse transcription; numbers indicate location of 5' nt

**Table 2.3 Species used in filtered multiple sequence alignment of HOTAIR to detect covariation.**

| Number in Alignment | Species ID | Species Name |
|---|---|---|
| 1 | HOTAIR_WT | *Homo sapiens* |
| 2 | CAUYPY01 | *Macaca cyclopis* |
| 3 | CAUYOM01 | *Trachypithecus pileatus* |
| 4 | CAUYPR01 | *Cercopithecus ascanius* |
| 5 | CAUYOP01 | *Allenopithecus nigroviridis* |
| 6 | CAUYNF01 | *Nomascus siki* |
| 7 | CAUYPQ01 | *Alouatta juara* |
| 8 | CAUYLE01 | *Cacajao ayresi* |
| 9 | CAUYLY01 | *Ateles paniscus* |
| 10 | CAUYMR01 | *Aotus azarai* |
| 11 | CAUYPS01 | *Leontopithecus chrysomelas* |
| 12 | CAUYOW01 | *Cebus unicolor* |
| 13 | CAUYTH01 | *Plecturocebus dubius* |
| 14 | CAUYQB01 | *Saguinus bicolor* |
| 15 | CAUYUG01 | *Leontocebus fuscicollis* |
| 16 | CAUYSM01 | *Saguinus oedipus* |
| 17 | CAUYOB01 | *Saimiri ustus* |
| 18 | CAUYNR01 | *Mico argentatus* |
| 19 | CAUYOR01 | *Callimico goeldii* |
| 20 | CAUYOE01 | *Alouatta macconnelli* |
| 21 | CAUYMP01 | *Eulemur sanfordi* |
| 22 | CAUYQI01 | *Hapalemur griseus alaotrensis* |
| 23 | CAUYQR01 | *Varecia variegata* |

| | | |
|---|---|---|
| 24 | CAUYSK01 | *Avahi laniger* |
| 25 | CAUYTA01 | *Propithecus verreauxi* |
| 26 | CAUYSO01 | *Propithecus diadema* |
| 27 | CAUYSF01 | *Eulemur flavifrons* |
| 28 | CAUYTU01 | *Indri indri* |
| 29 | CAUYSG01 | *Eulemur coronatus* |
| 30 | CAUYOY01 | *Cheirogaleus medius* |
| 31 | CAUYQH01 | *Nycticebus coucang* |
| 32 | CAUYQS01 | *Otolemur crassicaudatus* |
| 33 | CAUYRH01 | *Perodicticus potto* |
| 34 | CAUYQD01 | *Tarsius lariang* |
| 35 | CAUYQM01 | *Cephalopachus bancanus* |
| 36 | CAUYLF01 | *Cacajao hosomi* |
| 37 | CAUYMO01 | *Plecturocebus brunneus* |
| 38 | CAUYNH01 | *Ateles geoffroyi* |
| 39 | CAUYQJ01 | *Aotus griseimembra* |
| 40 | CAUYRM01 | *Cebus olivaceus* |
| 41 | CAUYNK01 | *Saimiri sciureus* |
| 42 | CAUYOQ01 | *Leontocebus fuscicollis* |
| 43 | CAUYQQ01 | *Mico humilis* |
| 44 | CAUYND01 | *Miopithecus ogouensis* |
| 45 | CAUYQO01 | *Nomascus gabriellae* |
| 46 | CAUYTB01 | *Gorilla beringei* |
| 47 | CAUYLW01 | *Varecia rubra* |
| 48 | CAUYNA01 | *Eulemur fulvus* |
| 49 | CAUYTD01 | *Hapalemur occidentalis* |
| 50 | CAUYTF01 | *Hapalemur griseus* |
| 51 | CAUYOC01 | *Propithecus tattersalli* |
| 52 | CAUYRB01 | *Propithecus perrieri* |
| 53 | CAUYMN01 | *Lepilemur ankaranensis* |
| 54 | CAUYON01 | *Arctocebus calabarensis* |
| 55 | CAUYRK01 | *Perodicticus potto ibeanus* |
| 56 | CAUYQV01 | *Galagoides demidoff* |
| 57 | CAUYTI01 | *Galago senegalensis* |
| 58 | CAUYPK01 | *Loris lydekkerianus* |
| 59 | CAUYQG01 | *Tarsius wallacei* |
| 60 | CAUYML01 | *Mirza zaza* |
| 61 | CAUYRT01 | *Lepilemur ruficaudatus* |
| 62 | CAUYPT01 | *Callithrix kuhlii* |
| 63 | CAUYTR01 | *Pithecia pissinatti* |
| 64 | CAUYOX01 | *Avahi peyrierasi* |
| 65 | CAUYPO01 | *Cheirogaleus major* |
| 66 | CAUYSU01 | *Cercopithecus nictitans* |

Species in a multiple sequence alignment of 66 primate genomes (including human HOTAIR) that was filtered by 95% sequence identity

**Table 2.4 Nucleotide variation in covariant base pairs in Domain 3 from a multiple sequence alignment of HOTAIR.**

| Left Position (nt) | Right Position (nt) | Base Pair (left:right) | Left nt | Sequence in Alignment | Right nt | Sequence in Alignment |
|---|---|---|---|---|---|---|
| 1367 | 1403 | C:G | C | 1-21, 23-25, 28-35, 51 | G | 1-33, 48 |
| | | | U | 36-40, 42-50, 52-53, 55-59, 64-66 | A | 36-47, 49-53, 55-59, 64, 66 |
| | | | | | C | 65 |
| 1428 | 1450 | A:U | A | 1-23, 27-35, 44, 66 | U | 1-23, 25-47, 49-53, 59, 64-66 |
| | | | U | 36 | G | 24, 48 |
| | | | G | 37-43, 45-53, 55-59, 64-65 | | |
| 1434 | 1444 | A:U | A | 1-33, 36-53, 59, 64-66 | U | 1-17, 19-25, 27-35, 47-53, 64-65 |
| | | | G | 34-35 | C | 26, 36-46, 55-59, 66 |
| | | | U | 55-58 | | |
| 1435 | 1443 | G:C | G | 1-25, 27-48, 50-53, 55-59, 64-66 | C | 1-31, 33-35, 49 |
| | | | A | 26 | U | 32, 36-48, 50-53, 55-59, 64-66 |
| 1436 | 1442 | G:C | G | 1-16, 18-33 | C | 1-25, 27-30, 34-35, 47-48, 50-53, 55-59, 64-65 |
| | | | A | 17, 34-35 | G | 26, 49 |
| | | | U | 36-48, 50-53, 55-59, 64-66 | U | 31-33 |
| | | | | | A | 36-46, 66 |
| 1437 | 1441 | U:G | U | 1-25, 27-48, 50-53, 55-58, 64-66 | G | 1-25, 27-34, 36-53, 55-59, 64-66 |
| | | | G | 26, 59 | A | 26, 35 |
| 1504 | 1521 | G:C | G | 1-6, 8-20 | C | 1-6, 8-20, 25, 28, 30-35, 52, 59 |
| | | | A | 7, 21-33, 35 | A | 21-23, 27, 29 |
| | | | C | 36-38, 40-53, 56, 59, 64-66 | G | 26, 36-50, 53, 55-58, 64-66 |
| | | | U | 55, 57-58 | | |
| 1505 | 1520 | C:G | C | 1-6, 8-19, 21-24, 27-33, 35-47, 49-53, 59, 64-66 | G | 1-6, 8-25, 27-32, 34, 48, 52-53, 59 |
| | | | G | 7, 48 | A | 26, 33, 36-47, 49-50, 55-57, 64-66 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | U | 20, 25, 55, 57-58 | U | 35 |
| | | | A | 26 | C | 58 |
| 1506 | 1517 | C:G | C | 1-6, 8-12, 14-20, 35-53, 55, 57-59, 64-66 | G | 1-6, 8-25, 27-35 |
| | | | U | 13, 21-23, 25-33 | A | 26, 58 |
| | | | G | 24 | C | 36-42, 44-51, 53, 55-57, 59, 64-66 |
| | | | | | U | 43, 52 |
| 1545 | 1556 | A:U | A | 1-6, 8-9, 11-20, 31-35 | U | 1-6, 8-25, 27-28, 30-49, 51-53, 55-58, 64-66 |
| | | | C | 10, 37 | C | 29 |
| | | | G | 21-25, 27-29, 36, 38-53, 55-57, 59, 64, 66 | G | 59 |
| | | | U | 65 | | |
| 1546 | 1555 | G:C | G | 1-6, 8-29, 31-32, 34-53, 55-57, 59, 64-66 | C | 1-2, 4-6, 8-28, 34-35 |
| | | | A | 33 | A | 3 |
| | | | | | U | 30, 36-39, 41-49, 51-53, 56-58, 65-66 |
| | | | | | G | 40, 59 |
| 1691 | 1703 | U:G | U | 1-6, 8-25, 27-29, 31-35 | G | 1-6, 9-12, 14-18, 20-25, 27-31, 36-41, 43-45, 51, 53, 55, 58, 64-66 |
| | | | C | 26, 36-46, 48, 50-53, 55, 58, 64-66 | A | 8, 13, 34, 50 |
| | | | A | 30 | U | 19, 26 |
| | | | | | C | 42 |
| 1692 | 1702 | U:A | U | 1-6, 8-46, 48, 50, 53, 55, 58, 64-66 | A | 1-6, 8, 10-25, 27, 29-31, 33-34, 36-40, 42-45, 48, 50-51, 53, 55, 58, 64-66 |
| | | | C | 51-52 | C | 9, 26 |
| | | | | | G | 28, 35, 41 |

Covariant base pairs were identified in a multiple sequence alignment of 66 primate genomes (including human HOTAIR) that was filtered by 95% sequence identity. Nucleotide number and base pair correspond to the position and identity in human HOTAIR. The sequence number corresponds to the "Number in Alignment" in **Table 2.3**. Sequences not included have gaps at that position.

## 2.9 References


1. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L. L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., Mercer, T. R., Moore, K. J., Nakagawa, S., Rinn, J. L., Spector, D. L., Ulitsky, I., Wan, Y., Wilusz, J. E. & Wu, M. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol*. **24**, 430-447.

2. Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. & Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*. **71**, 527-542.

3. Kopp, F. & Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*. **172**, 393-407.

4. Rinn, J. L. (2014). lncRNAs: linking RNA to chromatin. *Cold Spring Harb Perspect Biol*. **6**.

5. Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M. & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. **539**, 452-455.

6. Kawashima, H., Takano, H., Sugita, S., Takahara, Y., Sugimura, K. & Nakatani, T. (2003). A novel steroid receptor co-activator protein (SRAP) as an alternative form of steroid receptor RNA-activator gene: expression in prostate cancer cells and enhancement of androgen receptor activity. *Biochem J*. **369**, 163-171.

7. Li, X. L., Pongor, L., Tang, W., Das, S., Muys, B. R., Jones, M. F., Lazar, S. B., Dangelmaier, E. A., Hartford, C. C., Grammatikakis, I., Hao, Q., Sun, Q., Schetter, A., Martindale, J. L., Tang, B., Jenkins, L. M., Robles, A. I., Walker, R. L., Ambs, S., Chari, R., Shabalina, S. A., Gorospe, M., Hussain, S. P., Harris, C. C., Meltzer, P. S., Prasanth, K. V., Aladjem, M. I., Andresson, T. & Lal, A. (2020). A small protein encoded by a putative lncRNA regulates apoptosis and tumorigenicity in human colorectal cancer cells. *Elife*. **9**.

8. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. & Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. **129**, 1311-1323.

9. Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E. & Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science*. **329**, 689-693.

10. Portoso, M., Ragazzini, R., Brencic, Z., Moiani, A., Michaud, A., Vassilev, I., Wassef, M., Servant, N., Sargueil, B. & Margueron, R. (2017). PRC2 is dispensable for HOTAIR-mediated transcriptional repression. *EMBO J*. **36**, 981-994.

11. Brockdorff, N. (2013). Noncoding RNA and Polycomb recruitment. *RNA*. **19**, 429-442.

12. Meredith, E. K., Balas, M. M., Sindy, K., Haislop, K. & Johnson, A. M. (2016). An RNA matchmaker protein regulates the activity of the long noncoding RNA HOTAIR. *RNA*. **22**, 995-1010.

13. Porman, A. M., Roberts, J. T., Duncan, E. D., Chrupcala, M. L., Levine, A. A., Kennedy, M. A., Williams, M. M., Richer, J. K. & Johnson, A. M. (2022). A single N6-methyladenosine site regulates lncRNA HOTAIR function in breast cancer cells. *PLoS Biol*. **20**, e3001885.

14. Delhaye, L., De Bruycker, E., Volders, P. J., Fijalkowska, D., De Sutter, D., Degroeve, S., Martens, L., Mestdagh, P. & Eyckerman, S. (2022). Orthogonal proteomics methods to unravel the HOTAIR interactome. *Sci Rep*. **12**, 1513.

15. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S. & Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. **464**, 1071-1076.

16. Mozdarani, H., Ezzatizadeh, V. & Rahbar Parvaneh, R. (2020). The emerging role of the long non-coding RNA HOTAIR in breast cancer development and treatment. *J Transl Med*. **18**, 152.

17. Yang, Z., Zhou, L., Wu, L. M., Lai, M. C., Xie, H. Y., Zhang, F. & Zheng, S. S. (2011). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann Surg Oncol*. **18**, 1243-1250.

18. Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., Miyano, S. & Mori, M. (2011). Long noncoding RNA

HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res*. **71**, 6320-6326.

19. Loewen, G., Jayawickramarajah, J., Zhuo, Y. & Shan, B. (2014). Functions of lncRNA HOTAIR in lung cancer. *J Hematol Oncol*. **7**, 90.

20. Dong, L. & Hui, L. (2016). HOTAIR Promotes Proliferation, Migration, and Invasion of Ovarian Cancer SKOV3 Cells Through Regulating PIK3R3. *Med Sci Monit*. **22**, 325-331.

21. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. (2015). Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc*. **10**, 1643-1669.

22. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. & Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods*. **11**, 959-965.

23. Reuter, J. S. & Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*. **11**, 129.

24. Somarowthu, S., Legiewicz, M., Chillon, I., Marcia, M., Liu, F. & Pyle, A. M. (2015). HOTAIR forms an intricate and modular secondary structure. *Mol Cell*. **58**, 353-361.

25. Smola, M. J. & Weeks, K. M. (2018). In-cell RNA structure probing with SHAPE-MaP. *Nat Protoc*. **13**, 1181-1195.

26. Falese, J. P., McFadden, E. J., D'Inzeo, C. A. & Hargrove, A. E. (2025). Structural analysis of the lncRNA SChLAP1 reveals protein binding interfaces and a conformationally heterogenous retroviral insertion. *RNA*. **31**, 1260-1286.

27. Oh, M., Kadam, R. N., Charania, Z. S. & Somarowthu, S. (2025). LncRNA SChLAP1 Promotes Cancer Cell Proliferation and Invasion Via Its Distinct Structural Domains and Conserved Regions. *J Mol Biol*. **437**, 169350.

28. McCown, P. J., Wang, M. C., Jaeger, L. & Brown, J. A. (2019). Secondary Structural Model of Human MALAT1 Reveals Multiple Structure-Function Relationships. *Int J Mol Sci*. **20**.

29. Sherpa, C., Rausch, J. W. & Le Grice, S. F. (2018). Structural characterization of maternally expressed gene 3 RNA reveals conserved motifs and potential sites of interaction with polycomb repressive complex 2. *Nucleic Acids Res*. **46**, 10432-10447.

30. Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M. & Weeks, K. M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci U S A*. **113**, 10322-10327.

31. Marinus, T., Fessler, A. B., Ogle, C. A. & Incarnato, D. (2021). A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Res*. **49**, e34.

32. Busan, S. & Weeks, K. M. (2018). Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA*. **24**, 143-148.

33. Smola, M. J., Calabrese, J. M. & Weeks, K. M. (2015). Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry*. **54**, 6867-6875.

34. Wu, L., Murat, P., Matak-Vinkovic, D., Murrell, A. & Balasubramanian, S. (2013). Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. *Biochemistry*. **52**, 9519-9527.

35. Garbo, S., Minotti, S., Marocco, F., Quattrocchi, L., Di Silverio, I., Bontempi, G., Strippoli, R., Tripodi, M. & Battistelli, C. (2025). HOTAIR requires epitranscriptomic modification to exert its pivotal epigenetic role in Epithelial to Mesenchymal Transition. *Cell Death Dis*. **16**, 753.

36. He, S., Liu, S. & Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol*. **11**, 102.

37. Kuderna, L. F. K., Gao, H., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rousselle, M., Silva, F. E., Agueda, L., Blanc, J., Gut, M., de Vries, D., Goodhead, I., Harris, R. A., Raveendran, M., Jensen, A., Chuma, I. S., Horvath, J. E., Hvilsom, C., Juan, D., Frandsen, P., Schraiber, J. G., de Melo, F. R., Bertuol, F., Byrne, H., Sampaio, I., Farias, I., Valsecchi, J., Messias, M., da Silva, M. N. F., Trivedi, M., Rossi, R., Hrbek, T., Andriaholinirina, N., Rabarivola, C. J., Zaramody, A., Jolly, C. J., Phillips-Conroy, J., Wilkerson, G., Abee, C., Simmons, J. H., Fernandez-Duque, E., Kanthaswamy, S., Shiferaw, F., Wu, D., Zhou, L., Shao, Y., Zhang, G., Keyyu, J. D., Knauf, S., Le, M. D., Lizano, E., Merker, S., Navarro, A., Nadler, T., Khor, C. C., Lee, J., Tan, P., Lim, W. K., Kitchener, A. C., Zinner, D., Gut, I., Melin, A. D.,

Guschanski, K., Schierup, M. H., Beck, R. M. D., Umapathy, G., Roos, C., Boubli, J. P., Rogers, J., Farh, K. K. & Marques Bonet, T. (2023). A global catalog of whole-genome diversity from 233 primate species. *Science*. **380**, 906-913.

38. Tavares, R. C. A., Mahadeshwar, G., Wan, H. & Pyle, A. M. (2023). MRT-ModSeq - Rapid detection of RNA modifications with MarathonRT. *bioRxiv*.

39. Dai, Q., Zhang, L. S., Sun, H. L., Pajdzik, K., Yang, L., Ye, C., Ju, C. W., Liu, S., Wang, Y., Zheng, Z., Zhang, L., Harada, B. T., Dou, X., Irkliyenko, I., Feng, X., Zhang, W., Pan, T. & He, C. (2023). Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat Biotechnol*. **41**, 344-354.

40. Cavaco, M., Fraga, P., Valle, J., Andreu, D., Castanho, M. & Neves, V. (2021). Development of Breast Cancer Spheroids to Evaluate Cytotoxic Response to an Anticancer Peptide. *Pharmaceutics*. **13**.

41. Thermo Fisher Scientific, I. (2026). MDA-MB-231 Cell Line Spheroid Generation and Characterization for HT Assays. https://www.thermofisher.com/us/en/home/references/protocols/cell-culture/3-d-cell-culture-protocol/mda-mb-231-cell-line-spheroid-generation.html

42. Vinci, M., Box, C. & Eccles, S. A. (2015). Three-dimensional (3D) tumor spheroid invasion assay. *J Vis Exp*. e52686.

43. Yang, L., Liu, Y., Hu, X., Wang, P., Li, X. & Wu, J. (2021). Graph-Based Analysis of RNA Secondary Structure Similarity Comparison. *Complexity*.

44. Luo, D., Zheng, Y., Huang, Z., Wen, Z., Guo, L., Deng, Y., Li, Q., Bai, Y., Haider, S. & Wei, D. (2025). Exploiting functional regions in the viral RNA genome as druggable entities. *Elife*. **13**.

45. Finlay-Schultz, J., Jacobsen, B. M., Riley, D., Paul, K. V., Turner, S., Ferreira-Gonzalez, A., Harrell, J. C., Kabos, P. & Sartorius, C. A. (2020). New generation breast cancer cell lines developed from patient-derived xenografts. *Breast Cancer Res*. **22**, 68.

46. AAT Bioquest, I. (2026). RNA Molecular Weight Calculator. https://www.aatbio.com/tools/calculate-RNA-molecular-weight-mw

47. Guo, L. T., Adams, R. L., Wan, H., Huston, N. C., Potapova, O., Olson, S., Gallardo, C. M., Graveley, B. R., Torbett, B. E. & Pyle, A. M. (2020). Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J Mol Biol*. **432**, 3338-3352.

48. Slater, G. S. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. **6**, 31.

49. Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**, 772-780.

50. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150-3152.

51. Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658-1659.

52. Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G. & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol*. **39**.

53. Kuderna, L. F. K., Ulirsch, J. C., Rashid, S., Ameen, M., Sundaram, L., Hickey, G., Cox, A. J., Gao, H., Kumar, A., Aguet, F., Christmas, M. J., Clawson, H., Haeussler, M., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rouselle, M., Silva, F. E., Agueda, L., Blanc, J., Gut, M., de Vries, D., Goodhead, I., Harris, R. A., Raveendran, M., Jensen, A., Chuma, I. S., Horvath, J. E., Hvilsom, C., Juan, D., Frandsen, P., Schraiber, J. G., de Melo, F. R., Bertuol, F., Byrne, H., Sampaio, I., Farias, I., Valsecchi, J., Messias, M., da Silva, M. N. F., Trivedi, M., Rossi, R., Hrbek, T., Andriaholinirina, N., Rabarivola, C. J., Zaramody, A., Jolly, C. J., Phillips-Conroy, J., Wilkerson, G., Abee, C., Simmons, J. H., Fernandez-Duque, E., Kanthaswamy, S., Shiferaw, F., Wu, D., Zhou, L., Shao, Y., Zhang, G., Keyyu, J. D., Knauf, S., Le, M. D., Lizano, E., Merker, S., Navarro, A., Nadler, T., Khor, C. C., Lee, J., Tan, P., Lim, W. K., Kitchener, A. C., Zinner, D., Gut, I., Melin, A. D., Guschanski, K., Schierup, M. H., Beck, R. M. D., Karakikes, I., Wang, K. C., Umapathy, G., Roos, C., Boubli, J. P., Siepel, A., Kundaje, A., Paten, B., Lindblad-Toh, K., Rogers, J., Marques Bonet, T. & Farh, K. K. (2024). Identification of constrained sequence elements across 239 primate genomes. *Nature*. **625**, 735-742.

54. Tavares, R. C. A., Pyle, A. M. & Somarowthu, S. (2019). Phylogenetic Analysis with Improved Parameters Reveals Conservation in lncRNA Structures. *J Mol Biol*. **431**, 1592-1603.

55. Rivas, E., Clements, J. & Eddy, S. R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*. **36**, 3072-3076.

# 3. CHAPTER 3: RAPID AND ACCURATE RECONSTRUCTION OF FULL ATOMIC RNA STRUCTURES FROM COARSE-GRAINED MODELS

## 3.1 PREFACE

This work was completed under the mentorship of Dr. Chengxin Zhang and was a unique opportunity to develop computational structural biology software. Zion Perry developed the software, completed benchmarking, and drafted the manuscript. Dr. Chengxin Zhang conceptualized the project and developed the initial methodology and software. Dr. Chengxin Zhang and Dr. Anna Marie Pyle reviewed and edited the final manuscript. The Yale Center for Research Computing provided guidance and use of the high-performance computing clusters. This work was funded by the Howard Hughes Medical Institute (A.M.P.) and the National Science Foundation Graduate Research Fellowship Program (DGE-2139841 to Z.R.P.). A version of this chapter has been published in the *Journal of Molecular Biology* and is available online at https://doi.org/10.1016/j.jmb.2023.168210. Full citation: Perry, Z. R., Pyle, A. M., & Zhang, C. (2023). Arena: Rapid and Accurate Reconstruction of Full Atomic RNA Structures From Coarse-grained Models. *Journal of Molecular Biology*, 435(18), 168210.

## 3.2 ABSTRACT

RNA tertiary structures from experiments or computational predictions often contain missing atoms, which prevents analyses requiring full atomic structures. Current programs for RNA reconstruction can be slow, inaccurate, and/or require specific atoms to be present in the input. We present Arena (Atomic Reconstruction of RNA), which reconstructs a full atomic RNA structure from residues that can have as few as one atom. Arena first fills in missing atoms and then iteratively refines their placement to reduce nonideal geometries. We benchmarked Arena on

a dataset of 361 RNA structures, where Arena achieves high accuracy and speed compared to other structure reconstruction programs. For example, Arena was used to reconstruct full atomic structures from a single phosphorus atom per nucleotide to, on average, within 3.63 Å RMSD of the experimental structure, while virtually removing all clashes and running in < 3 sec, which is 353× and 46× faster than state-of-the-art programs PDBFixer and C2A, respectively. The Arena source code and webserver are freely available at https://github.com/pylelab/Arena and https://zhanggroup.org/Arena/, respectively.

## 3.3 Introduction

Elaborate three-dimensional (3D) structures enable RNA to participate in diverse cellular processes. Developments in high-resolution cryo-electron microscopy (cryo-EM) for RNA 3D structure determination make this an exciting time for studying RNA structure-based mechanisms [1]. Full atomic structures provide the most insight into structure-function relationships of RNA molecules, such as better understanding RNA folding and catalysis [2], designing small molecule drugs [3], and performing more accurate computational simulations [4].

Experimental and computational studies often generate partially or fully coarse-grained models, which lack full atomic information. In RNA crystallography and cryo-EM, often the phosphate and base atoms are the most well-resolved, but much of the structure lacks sufficient resolution to determine all the atoms [5]. Computational programs for RNA 3D structure prediction, such as NAST [6] and SimRNA [7], improve simulation speed using coarse-grained representations, which require other post-processing methods to reconstruct a full atomic model. The challenge of full atomic reconstruction from coarse-grained models has largely been solved for proteins with PULCHRA [8], which reconstructs full atomic structures quickly and accurately.

The programs available for RNA full atomic reconstruction have begun to address this challenge, but improvements in accuracy, speed, and ease of use are needed.

State-of-the-art programs that most directly solve this challenge for RNA are PDBFixer [9] and C2A [10]. PDBFixer adds missing atoms to a target structure using a standard template PDB for each residue and superimposes the template atoms onto the target atoms. PDBFixer then runs a set of short Langevin dynamics simulations to reduce clashes. C2A requires a coarse-grained target, a definition of RNA secondary structure segments, and a reference full atomic 3D structure (the default is PDB 1N32) from which full atomic structure fragments are derived. C2A chooses fragments in the reference structure that match the target on a coarse-grained level using a Monte Carlo simulation to reduce collisions and then builds the reference fragment's atomic detail into the target.

Two other programs reconstruct a subset of atoms. Rosetta rna_thread [11], originally created for homology modeling, can reconstruct side chains given the sugar-phosphate backbone. RCrane [5], originally created for RNA crystallography refinement, reconstructs the sugar-phosphate backbone given the base, phosphorus (P), and C1' atoms. RCrane performs backbone reconstruction by matching a library of empirically determined backbone conformers [12] with the backbone pseudo-torsion angles derived from the base, P, and C1' atom positions. The sugar conformation is further refined by downhill simplex minimization. We used RCrane.CLI, which is a fully automated version of RCrane that builds the structure from the given atom positions without the need to locate them in an electron density map.

Here, we introduce Arena (Atomic Reconstruction of RNA), which is highly accurate, fast, and easy to use. Notably, Arena generates a full atomic structure for any coarse-grained RNA which has at least one atom present per nucleotide, regardless of its identity. Arena first

superimposes atoms from an ideal RNA helix [13] onto the target and iteratively refines atom placement by correcting bond lengths, bond angles, base and base pair conformations, and steric clashes. Arena, written in C++, is fully automated, and does not require the user to install third-party packages or libraries nor provide additional inputs, such as sequences, structural templates, or experimental data. The program is freely available at https://github.com/pylelab/Arena and as a webserver at https://zhanggroup.org/Arena/.

## 3.4 RESULTS

### *3.4.1 Overview of Arena*

Arena (Atomic Reconstruction of RNA) takes an RNA structure with missing atoms and produces a full atomic, refined structure (**Figure 3.1**). As input, Arena takes an RNA in PDB format, which can be of any length and must contain at least one atom per nucleotide of any identity (**Figure 3.1A**). Arena first fills in all the missing atoms for each nucleotide by superimposing fragments from an ideal A-form RNA helix [13] onto the input structure (**Figure 3.1B**). Meanwhile, the secondary structure is assigned with CSSR [14] prior to refinement (**Figure 3.1C**). The structure then undergoes multiple steps of refinement to correct nonideal bond lengths and angles (**Figure 3.1D**), correct base and base pair conformations (**Figure 3.1E**), and remove clashing atom pairs (**Figure 3.1F**). Arena iterates through the refinement steps according to the length of the RNA (between 100-1,000 iterations) and updates the secondary structure assignment every $10^{th}$ iteration (**Figure 3.1C**). This results in a refined, full atomic structure in < 3.5 sec on average (**Figure 3.1G**).

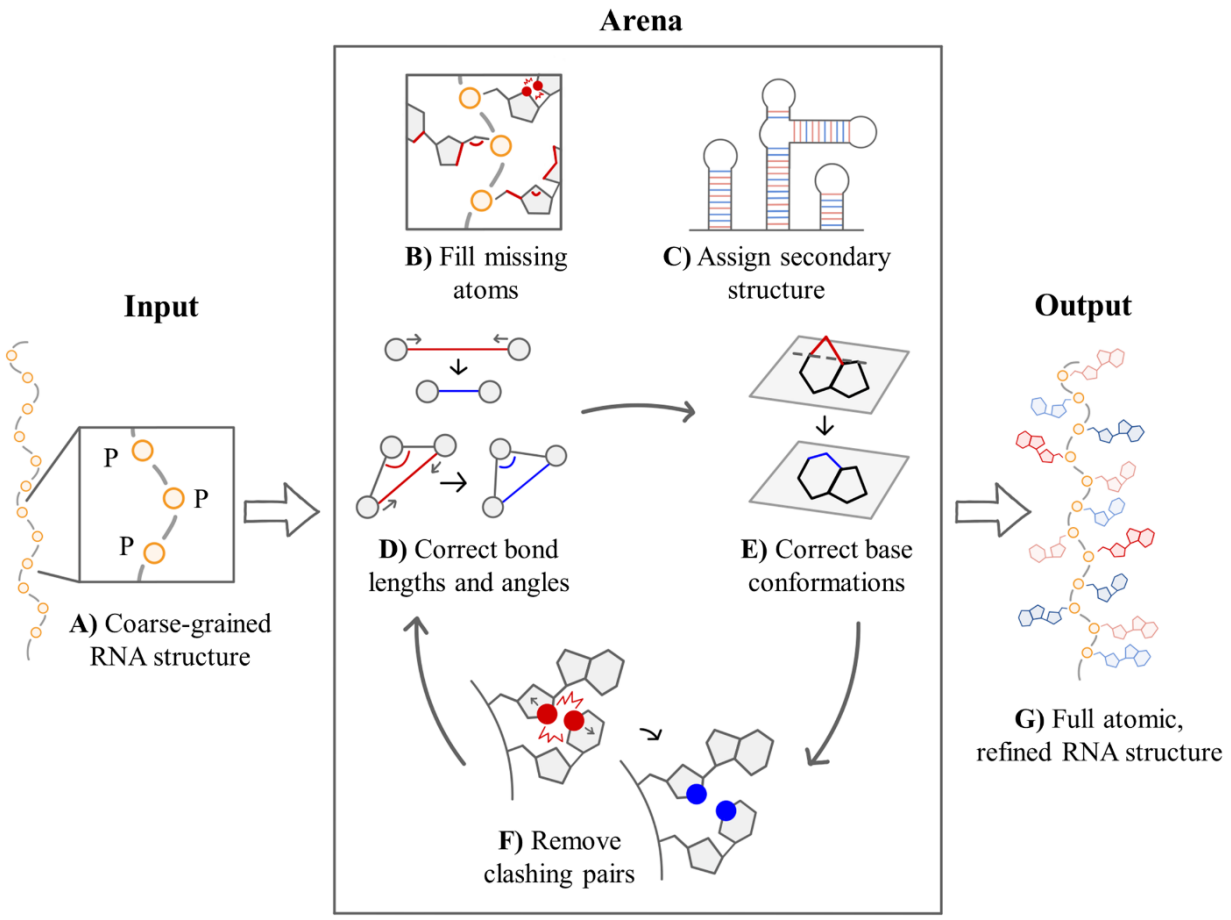



**Figure 3.1 Computational workflow of Arena.**
Starting from a coarse-grained RNA structure, shown here containing only phosphorus (P) atoms (A), Arena fills in any missing non-hydrogen atoms (B) and assigns secondary structure (C). Arena then iterates through several stages of refinement to correct bond lengths and bond angles (D), correct base and base pair conformations (E), and remove clashing atomic pairs (F) to generate a fully atomic and refined RNA structure (G).

### *3.4.2 Benchmarking Datasets*

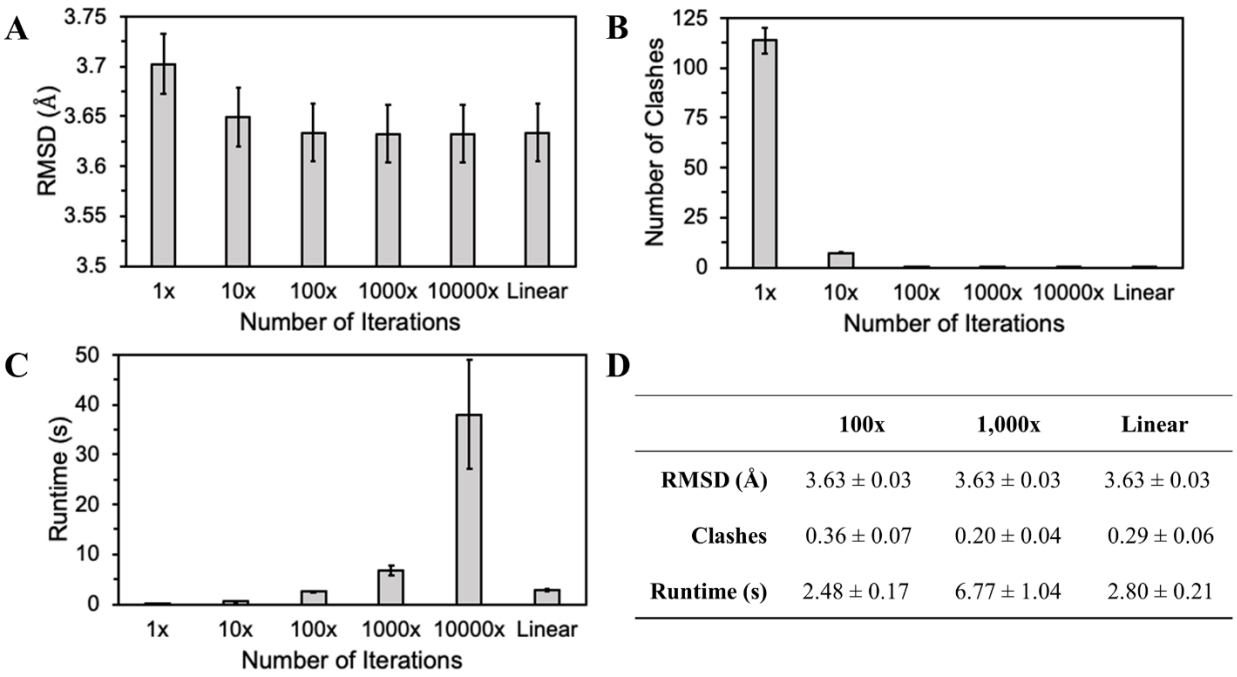
Arena and other programs were benchmarked on a dataset of 361 non-redundant RNA structures curated from the PDB according to previous studies [14, 15]. Briefly, the RNAs, which include rRNAs, tRNAs, snRNAs, introns, riboswitches, etc., are single chains between 30 and 692 nucleotides, have ≥ 10 canonical base pairs, have resolution < 4 Å, and share < 80% sequence identity.

From this PDB dataset, we generated 5 datasets to evaluate the ability of Arena and other programs to rebuild RNAs from different starting structures and/or due to the input requirements of other programs. Each dataset contains a different subset of atoms, including 1) the backbone atoms (P, OP1, OP2, O5', C5', C4', O4', C3', O3', C2', O2', C1'), 2) the P, C1', and side chain atoms (N9, C8, N7, C5, C6, N6, N1, C4, C2, O2, N3, N4, O6, N2, O4), 3) the P atoms, 4) the C3' atoms, and 5) the N atoms involved in the N-glycosidic bond (the N1 atom for nucleotides C and U or the N9 atom for nucleotides A and G). Additionally, we generated 2 computationally predicted coarse-grained datasets from NAST (C3' atoms) and SimRNA (P, C4', N9, C2 and C6 for nucleotides G and A; P, C4' N1, C2 and C4 for nucleotides C1 and U) (See **3.6 Materials and Methods**). All PDB files and source code for benchmarking are available at https://zenodo.org/records/18963141.

We evaluated each program's performance with four metrics, the heavy atom root-mean-square-deviation (RMSD) from the experimental structure, interaction network fidelity (INF), number of atom pairs that clash with each other, and computational runtime. INF quantifies the consistency of base-base interactions between two structures, where an INF score of 1 indicates perfect agreement (See **3.6 Materials and Methods**) [16]. While Arena, PDBFixer, rna_thread, and RCrane keep the coordinates of the original input atoms unchanged in the output structure, C2A may shift the coordinates of the input atoms. Since this coordinate shift is caused by superimposition of local fragments onto the input structure, the C2A output is in the same frame as the input structure. Therefore, there is no need to perform global superimposition of the output structure onto the native structure before the RMSD calculation. To obtain consistent runtime statistics, all benchmarks were performed single-threaded on Intel Xeon E5-2660 v3 CPUs.

*3.4.3 Optimization of Arena*

We first optimized the number of iterations that Arena performs to correct bond lengths and angles, base and base pair conformations, and clashes (**Figure 3.1D-F**). We tested between 1 to 10,000 iterations and used these results to inform a linear function where the number of iterations is proportional to the length of the input RNA, with a minimum of 100 iterations (**Figure 3.2**). This function provides a good balance between accuracy (RMSD of 3.63 Å, 0.29 clashes) and runtime (2.80 s). Performing 10,000 iterations sharply increases the average runtime (38.09 s ± 10.99), most severely for longer sequences, while the accuracy is not significantly improved (RMSD of 3.63 Å ± 0.03, 0.21 clashes ± 0.05). Thus, between 100-1,000 iterations is satisfactory for all input RNAs tested.




| | 100x | 1,000x | Linear |
|---|---|---|---|
| **RMSD (Å)** | 3.63 ± 0.03 | 3.63 ± 0.03 | 3.63 ± 0.03 |
| **Clashes** | 0.36 ± 0.07 | 0.20 ± 0.04 | 0.29 ± 0.06 |
| **Runtime (s)** | 2.48 ± 0.17 | 6.77 ± 1.04 | 2.80 ± 0.21 |

**Figure 3.2 Optimization of the number of iterations of Arena.**
Multiple iterations of the refinement stages (between 1-10,000 iterations) were tested prior to benchmarking and evaluated according to RMSD (A), clashes (B), and runtime (C), with average values summarized (D). Linear indicates the function where the number of iterations is proportional to the length *L* of the RNA (*0.4*L + 100*).

*3.4.4 Performance of Arena*

We evaluated the performance of Arena on the datasets described above. First, we evaluated reconstruction of the side chains from structures containing the backbone atoms (**Table 3.1, Figure 3.3A-D**). Arena outperformed the other programs with an RMSD of 0.81 Å, INF of 0.83, 30.66 clashes, and runtime of 3.21 s. Rosetta rna_thread had the next best RMSD (0.98 Å), INF (0.81), and runtime (4.43 s) after Arena but had on average 12× more clashes (374.21 clashes).

**Table 3.1 Performance of programs on reconstruction from backbone atoms.**

| | Arena | PDBFixer | C2A | rna_thread |
|---|---|---|---|---|
| **RMSD (Å)** | 0.81 ± 0.01* | 2.25 ± 0.02 (4.42E-221) | 4.70 ± 0.16 (1.21E-82) | 0.98 ± 0.01 (1.89E-67) |
| **INF** | 0.83 ± 0.00* | 0.09 ± 0.00 (0) | 0.56 ± 0.01 (6.24E-193) | 0.81 ± 0.00 (5.70E-21) |
| **Clashes** | 30.66 ± 2.10* | 221.93 ± 10.86 (3.65E-61) | 586.81 ± 37.14 (6.34E-42) | 374.21 ± 18.98 (1.68E-59) |
| **Runtime (s)** | 3.21 ± 0.23* | 485.11 ± 76.20 (6.81E-10) | 222.22 ± 59.76 (2.75E-4) | 4.43 ± 0.08 (1.15E-11) |
| **MAE (Å)** | 0.17 ± 0.00* | 0.59 ± 0.01 (1.26E-209) | 1.66 ± 0.06 (2.87E-84) | 0.19 ± 0.00 (1.70E-28) |

Average values are reported with standard error of the mean. *p-value between Arena and each program was calculated using a paired, two-tailed t-test.

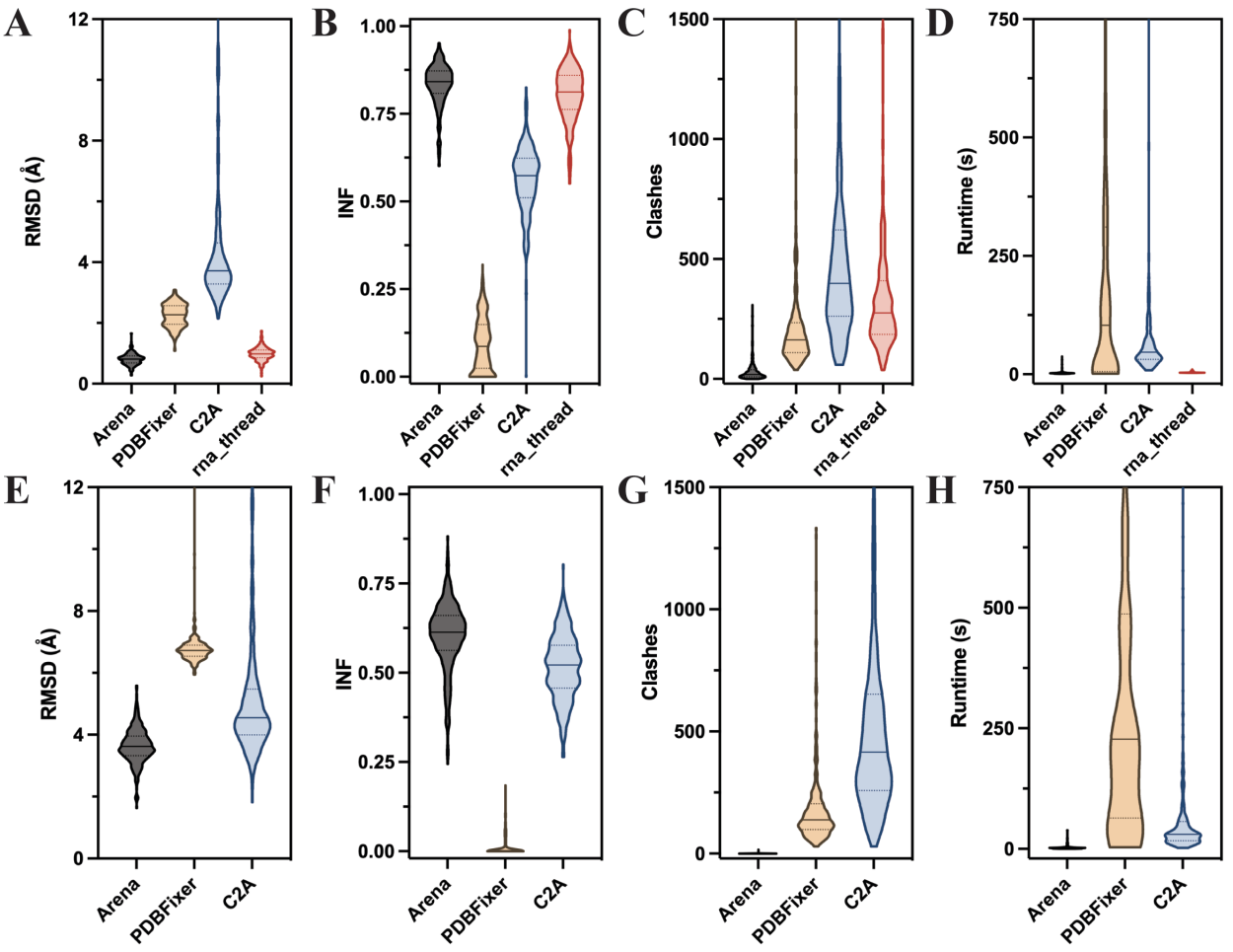



**Figure 3.3 Performance of programs on reconstruction from backbone atoms and P atoms.** Arena, PDBFixer, C2A, and rna_thread were used to reconstruct the side chains from input containing the backbone atoms, for which RMSD (A), INF (B), clashes (C), and runtime (D) were calculated. Arena, PDBFixer, and C2A were then used to reconstruct the full structure from input only containing the P atoms, for which RMSD (E), INF (F), clashes (G), and runtime (H) were calculated. The violin plots show the median (solid line) and quartiles (dotted lines) and were created in GraphPad Prism.

Next, we evaluated reconstruction of the backbone from structures containing the P, C1', and base atoms (**Table 3.5, Figure 3.4A-D**). While Arena is 70× faster and contains 77% fewer clashes than the next best program (RCrane), the RMSD of RCrane is on average 0.11 Å lower than that of Arena. Since all the base atoms were defined by the input, the INF score is 1.00 except for C2A, which moves the input atom positions. Whereas Arena does not consider backbone

torsion angles, RCrane uses a pseudotorsional system that is aided by a library of empirically determined backbone conformers [12]. Although Arena allows flexibility of the phosphate and ribose groups, the final structure may be biased by the initial backbone conformation used by Arena, which is fixed by the ideal RNA helix template. Conversely, RCrane adjusts the initial conformation based on the coordinates of the P and base atoms, which explains why RCrane has a slightly better RMSD for backbone reconstruction than Arena.

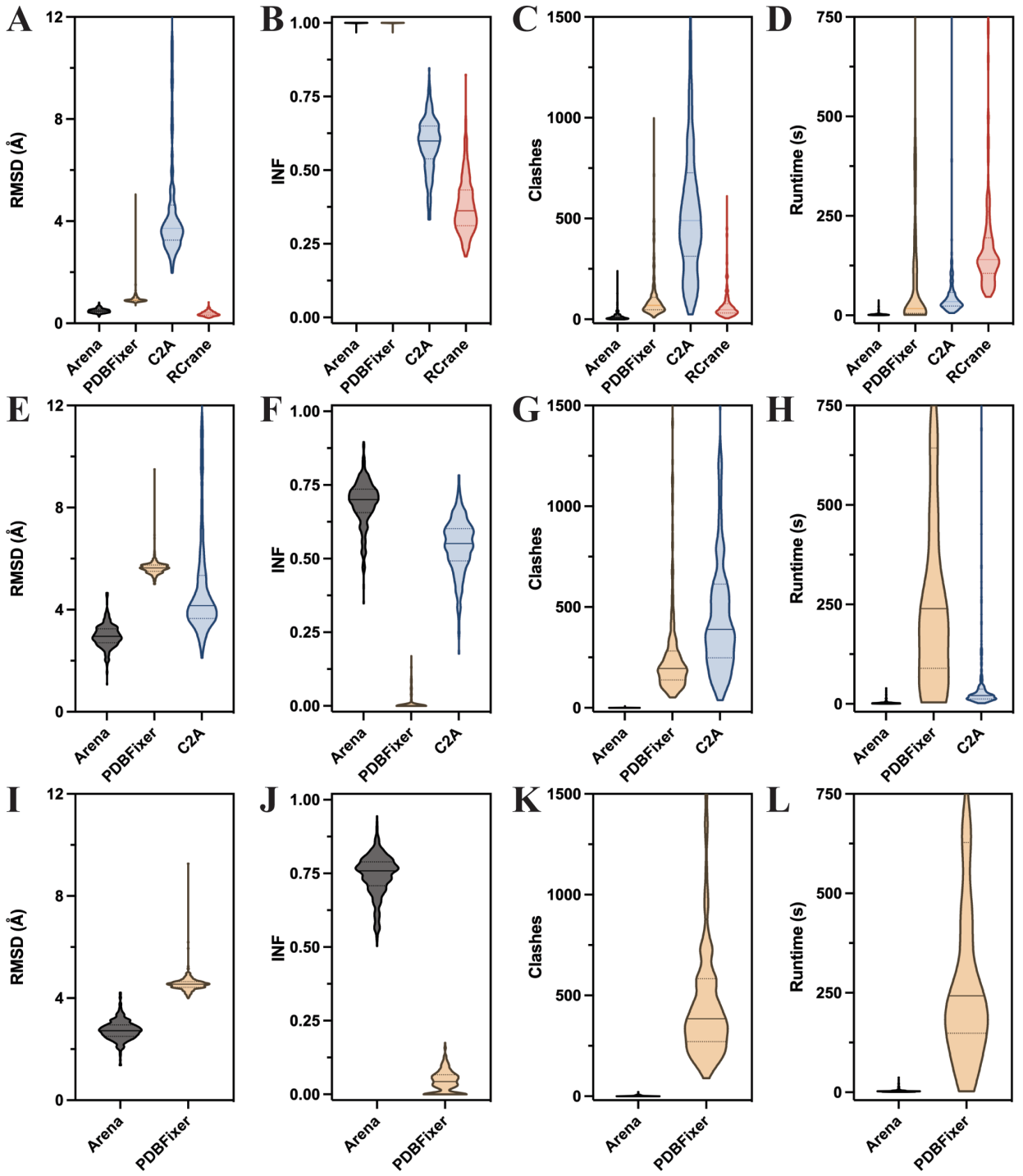



**Figure 3.4 Performance of programs on reconstruction from P, C1', and base atoms, from C3' atoms, and from glycosidic N atoms.**

Arena, PDBFixer, C2A, and RCrane.CLI were used to reconstruct the backbone from input containing the P, C1', and base atoms, for which RMSD (A), INF (B), clashes (C), and runtime (D) were calculated. Arena, PDBFixer, and C2A were used to reconstruct the full structure from input only containing the C3' atoms, for which RMSD (E), INF (F), clashes (G), and runtime (H) were calculated. Arena and PDBFixer were used to reconstruct the full structure from input only

containing the glycosidic N atoms, for which RMSD (I), INF (J), clashes (K), and runtime (L) were calculated. The violin plots show the median (solid line) and quartiles (dotted lines) and were created in GraphPad Prism.

We next tested the more challenging tasks of reconstruction from input models of only one atom per nucleotide. We started with models containing only the P atom, which is the most relevant usage case of Arena for experimental biologists, as ~5.6% of RNA chains in the PDB database only contain P atoms [14]. Arena significantly outperformed PDBFixer and C2A with an RMSD of 3.63 Å, INF of 0.60, 0.28 clashes, and runtime of 2.70 s (**Table 3.2, Figure 3.3E-H**). We then evaluated reconstruction from input containing only the C3' atoms, which is a representation used by the NAST and NASTI programs for RNA structures [17] (**Table 3.6, Figure 3.4E-H**). Arena outperformed PDBFixer and C2A with an RMSD of 2.98 Å, INF of 0.69, 0.23 clashes, and runtime of 2.93 s. Lastly, we used structures with only the glycosidic N atoms to evaluate Arena's reconstruction from a single side chain atom and completely undefined backbone. Arena again outperformed PDBFixer with a 60% lower RMSD (2.74 Å), 19x higher INF (0.74), 99.6% fewer clashes (2.24 clashes), and a 275× faster runtime (3.48 s) (**Table 3.7, Figure 3.4I-L**).

**Table 3.2 Performance of programs on reconstruction from P atoms.**

| | Arena | PDBFixer | C2A |
|---|---|---|---|
| **RMSD (Å)** | 3.63 ± 0.03* | 6.77 ± 0.03<br>(7.10E-228) | 5.15 ± 0.11<br>(4.32E-37) |
| **INF** | 0.60 ± 0.01* | 0.01 ± 0.00<br>(2.93E-285) | 0.52 ± 0.00<br>(7.21E-49) |
| **Clashes** | 0.28 ± 0.06* | 186.73 ± 9.27<br>(4.13E-61) | 635.04 ± 42.50<br>(1.22E-39) |
| **Runtime (s)** | 2.70 ± 0.23* | 953.50 ± 145.46<br>(2.04E-10) | 122.93 ± 24.33<br>(9.98E-7) |

| | | | |
|---|---|---|---|
| **MAE (Å)** | 1.13 ± 0.01* | 2.66 ± 0.01 (3.62E-283) | 1.85 ± 0.05 (3.70E-49) |

Average values are reported with standard error of the mean. *p-value between Arena and each program was calculated using a paired, two-tailed t-test.

Next, we tested each program's ability to rebuild structures from computationally predicted coarse-grained structures using NAST and SimRNA (See **3.6 Materials and Methods**). Both programs predict the RNA's 3D structure from sequence and generate a coarse-grained structure of C3' atoms (NAST) or the P, C4', glycosidic N, C2 and C4 atoms (SimRNA). We then compared Arena's performance to PDBFixer and the corresponding full atomic reconstruction program written for NAST (C2A) or SimRNA. Arena outperformed both PDBFixer and C2A for the NAST-generated inputs (**Figure 3.5A-D, Table 3.8**) and outperformed or was on par with PDBFixer and SimRNA for the SimRNA-generated inputs (**Figure 3.5E-G, Table 3.9**). However, the initial model accuracy will affect the final reconstruction. For example, reconstruction from the experimentally determined C3' positions (**Table 3.6**) results in a lower RMSD, higher INF, and lower clashes compared to reconstruction from the computationally determined C3' positions (**Table 3.8**). Moreover, for 11.1% of SimRNA structures, Arena-reconstructed models have at least one backbone discontinuity, which was mainly caused by unusual local conformations of the SimRNA coarse-grained models (See Figure 3.6 and Table 3.10 for an example on PDB 6Q9A, chain 4). To address this issue, we benchmarked Arena (extended) on the SimRNA dataset, which performs an additional simulation after the default run where all atoms are allowed to move. This fixes any discontinuous structures and removes 99.9% of clashes for a slight increase in runtime by approximately 0.2 s (**Figure 3.5E-H, Table 3.9**). We also benchmarked Arena (extended) on the NAST dataset, which removes 92.5% of clashes for an approximately 0.9 s increase in runtime (**Figure 3.5A-D, Table 3.8**). Overall, for the coarse-grained predictions from NAST and SimRNA,

Arena can generate a full atomic structure at least as good as, if not better than, the full atomic reconstruction program developed specifically for each coarse-grained prediction program.

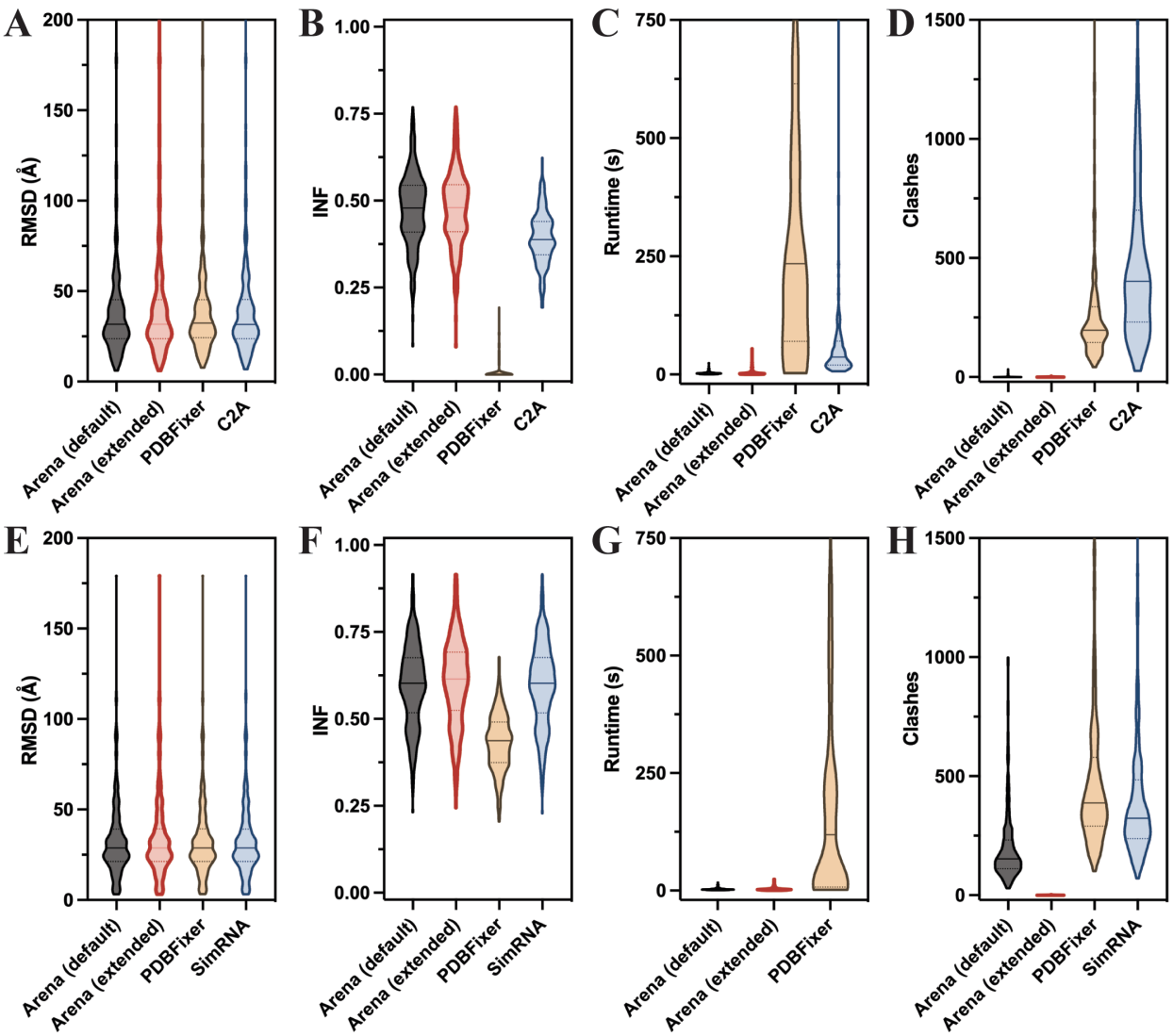



**Figure 3.5 Performance of programs on reconstruction from coarse-grained structure predictions.**
Arena (default and extended options), PDBFixer, and C2A were used to reconstruct the full atomic structure from coarse-grain models predicted by NAST, for which RMSD (A), INF (B), clashes (C), and runtime (D) were calculated. Arena (default and extended options), PDBFixer, and SimRNA's internal module were used to reconstruct the full atomic structure from coarse-grain models predicted by SimRNA, for which RMSD (E), INF (F), clashes (G), and runtime (H) were calculated. The violin plots show the median (solid line) and quartiles (dotted lines) and were created in GraphPad Prism.

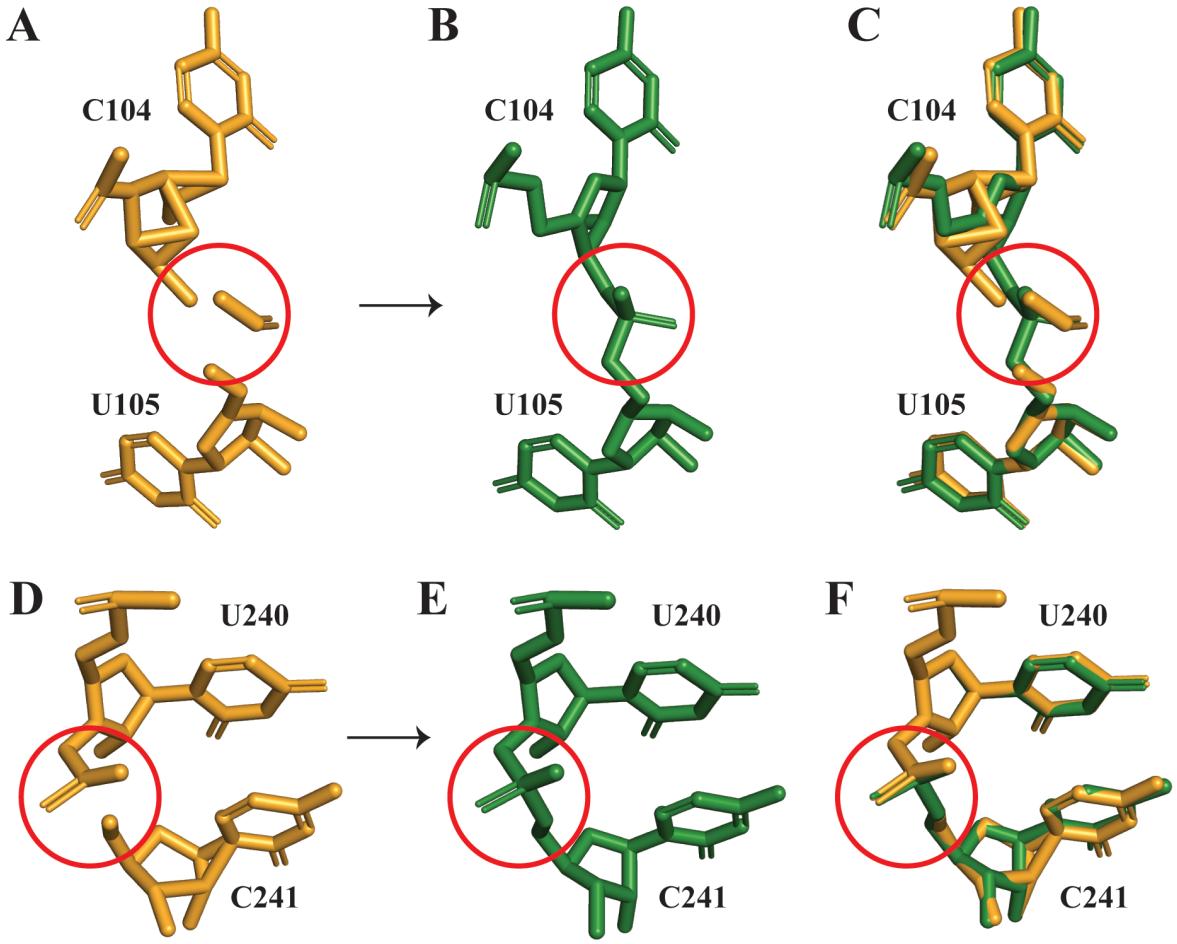



**Figure 3.6 The extended Arena simulation closes discontinuous nucleotides in a SimRNA-generated structure.**
Arena (default) and Arena (extended) were applied to build the full atom structure of the SimRNA-predicted model of a 363-nt tmRNA (PDB 6Q9A, chain 4). The Arena (default) simulation resulted in gaps (red circle) between nucleotides 104-105 (A) and 240-241 (D), which were fixed by the Arena (extended) simulation (B, E). The output structure from the default (orange) and extended (green) simulation are overlaid (C, F). Structures were visualized in PyMOL.

Arena can also be used to rebuild RNA models of NMR structures, and we have provided an additional script for this use case. We tested Arena on two NMR structures (2N3Q and 2KOC) using only the P atoms as input (**Table 3.11**) and find that Arena's performance is similar to that on single-model structures (**Table 3.2**). Finally, we compared Arena's performance on different classes of RNA, including tRNAs, rRNAs, snRNAs, ribozymes, and riboswitches (**Figure 3.7,**

**Table 3.12**). Compared to the full benchmarking dataset, Arena performs slightly better on ribozymes and riboswitches and slightly worse on rRNAs and snRNAs, while there is no difference in performance on tRNAs (significance threshold of $p \leq 0.05$). While Arena can satisfactorily reconstruct any input RNA, it may perform better on shorter and/or more structurally defined RNAs, such as some ribozymes and riboswitches, which may have more accurate initial atom positions.

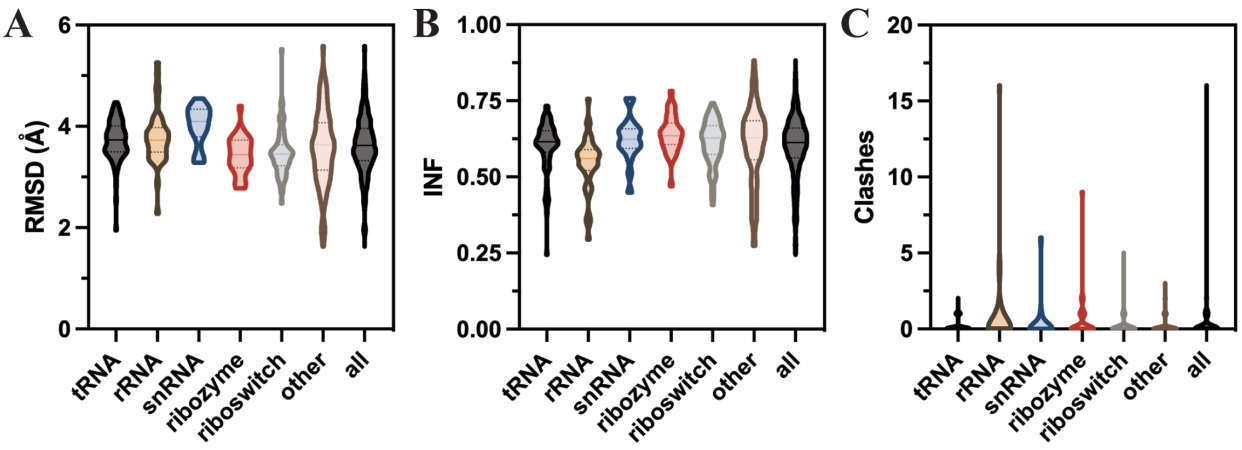



**Figure 3.7 Comparison of Arena's performance on different classes of RNA.**
RMSD (A), INF (B), and clashes (C) are shown for different types of RNA (tRNA, rRNA, snRNA, ribozyme, riboswitch, other) from benchmarking Arena on the input dataset containing only P atoms. The label "all" indicates the full dataset of 361 structures generated from P atoms. The violin plots show the median (solid line) and quartiles (dotted lines) and were created in GraphPad Prism.

We report the Pearson correlation coefficients between performance metrics for the reconstruction from P atoms by Arena (**Figure 3.8**). As expected, the runtime increases linearly with the length ($r = 0.94$) (**Figure 3.8D**). Although the absolute number of clashes increases with the length ($r = 0.45$) (**Figure 3.8C**), larger structures do not necessarily have more clashes when clashes are normalized by the length ($r = 0.14$) (**Figure 3.8E**). A lower RMSD generally tracks with a higher INF ($r = -0.66$) (**Figure 3.8F**), although the two metrics are independently valuable.

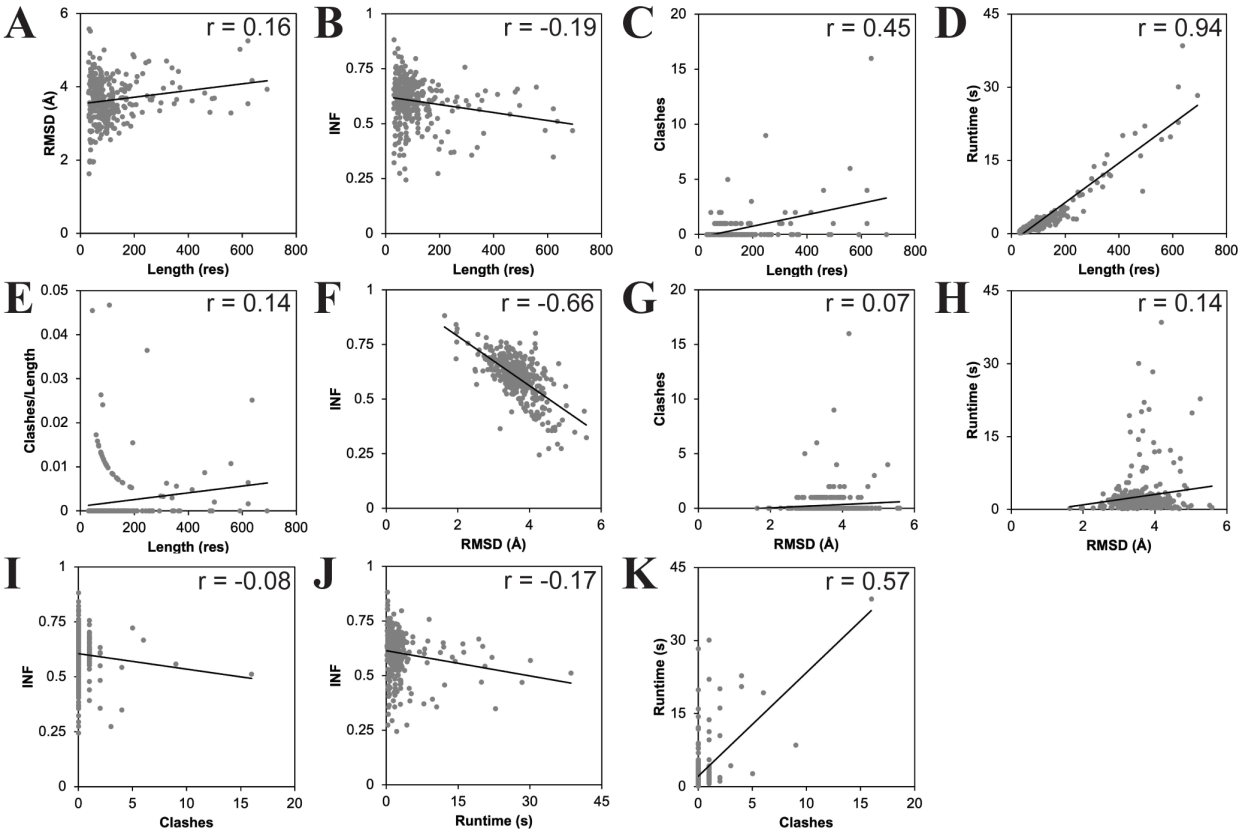



**Figure 3.8 Relationships between performance metrics.**
The results of benchmarking Arena on the input dataset containing only P atoms were plotted to check for correlations between each pair of performance metrics. The plots show the linear trendline and Pearson correlation coefficient (r).

There are several reasons why Arena is much faster than the other programs tested. Firstly, Arena does not perform any stochastic Monte Carlo or molecular dynamics simulations but instead uses an ultra-fast deterministic simulation that performs distance-based calculations. Arena thus has fewer degrees of freedom when searching through conformational space, so it more quickly generates a final structure. More importantly, since all movements in the deterministic simulation are based on ideal RNA geometries, the simulation effectively avoids sampling of suboptimal conformations, thereby producing more accurate conformations than stochastic simulations. Arena also performs a defined number of iterations regardless of performance. By contrast, PDBFixer performs clash removal using the OpenMM package [9], whose simulation may become locally

trapped if clashes cannot be readily removed. Finally, Arena is written in highly compact C++ code, which is much faster than the scripting languages used by the other programs, such as PDBFixer (Python 3), C2A (Python 2.5), and RCrane (Perl).

### *3.4.5 Case Studies on Two RNA Structures*

Notably, Arena is able to reduce or completely remove all clashing atom pairs. We demonstrate this with the reconstruction of an 88-nt 18S rRNA fragment (PDB 5OSG, chain 2) from an input containing only the P atoms (**Figure 3.9, Table 3.3**). Compared to PDBFixer and C2A, Arena achieves a lower-RMSD structure with INF of 0.54 in less than 6% of the runtime (**Figure 3.9A**). The PDBFixer-generated structure fails to generate any base-base interactions (**Figure 3.9B, Figure 3.9E**), while the C2A-generated structure, although reproducing base-base interactions like Arena (INF of 0.55), has large structural shifts in the backbone compared to the experimental structure (**Figure 3.9C**). PDBFixer and C2A leave 209 and 592 clashes, respectively, while Arena removes all clashes in the final structure. The number of clashes for each structure is consistent with the average number of clashes from the full-scale benchmarking on P atom structures (averages of 0.28, 186.73, and 635.04 clashes for Arena, PDBFixer, and C2A, respectively). An example region with clashes is shown with residues 2-4 and 11-13 for Arena (**Figure 3.9D**), PDBFixer (**Figure 3.9E**), and C2A (**Figure 3.9F**). We also noticed that PDBFixer sometimes creates nonbiological bonds, such as in the ribose ring of residue G12 (**Figure 3.9E** inset).

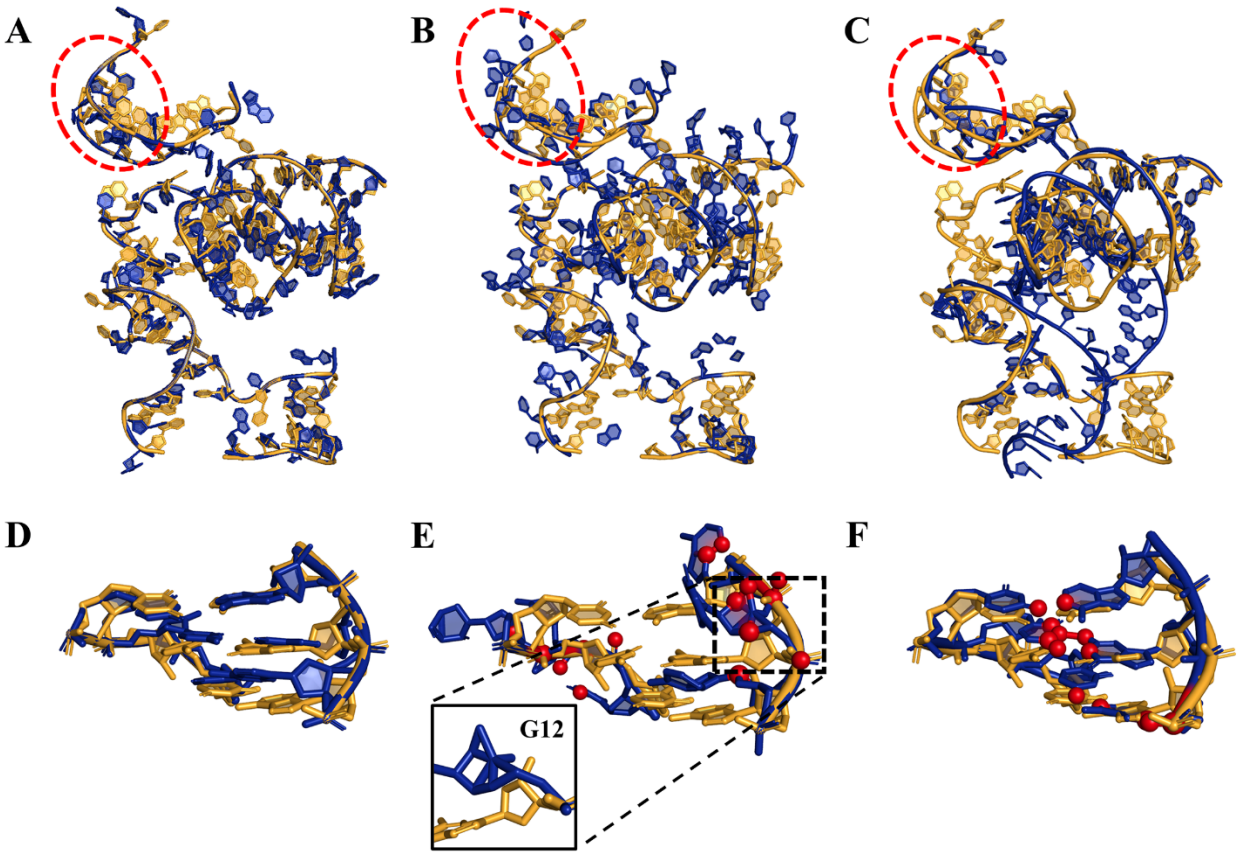



**Figure 3.9 Arena generates a structure close to native with no clashes.**
The native structure (orange) of the 88-nt 18S rRNA fragment (PDB 5OSG, chain 2) is overlaid with the output structure (blue) generated from an input containing only the P atoms by the programs Arena (A, D), PDBFixer (B, E), and C2A (C, F). The red oval approximately indicates residues shown in D-F (res 2-4, 11-13). Atoms that are part of one or more clashing atom pairs are shown as red spheres. The nonbiological ribose ring of residue G12 that was generated by PDBFixer (blue) is shown compared to the native ring (orange) (E, inset). Structures were visualized in PyMOL.

**Table 3.3 Performance of programs on reconstruction of 5OSG from P atoms.**

| | Arena | PDBFixer | C2A |
|---|---|---|---|
| **RMSD (Å)** | 4.15 | 7.72 | 9.52 |
| **INF** | 0.54 | - | 0.55 |
| **Clashes** | 0 | 209 | 592 |
| **Runtime (s)** | 1.64 | 289.71 | 28.92 |

| | | | |
|---|---|---|---|
| **MAE (Å)** | 1.40 | 2.82 | 3.39 |

A user may commonly encounter experimental structures that are partially or fully coarse-grained, which accounts for at least 5.6% of RNAs deposited in the PDB [14]. In such cases, users may still need the corresponding full atomic models, which are required for many downstream analyses such as detection of non-canonical base pairs or discovery of atomic contacts between ligands and RNAs. Therefore, we tested Arena's ability to reconstruct a partially coarse-grained structure outside the benchmarking dataset. We chose the cryo-EM structure of the 146-nt 9S rRNA fragment (PDB 6SG9, chain CA) (**Figure 3.10, Table 3.4**). Part of the structure is full atomic, but there are large stretches of residues missing side chains. We used each program to generate a full atomic structure, which we compared to the experimental electron density map. We calculated the average (**Table 3.4**) and per-residue (**Figure 3.11**) Q-scores between each output and the electron density map, where a higher Q-score means a greater agreement between the atomic positions and the density map [18]. Arena results in the same Q-score as the original structure (0.34) (**Figure 3.10A**), compared to 0.33 for PDBFixer (**Figure 3.10B**), 0.04 for C2A due to large shifts in the backbone (**Figure 3.10C**), and 0.33 for rna_thread (**Figure 3.10D**). This shows that Arena successfully adds more atomic details without ruining the original structure.

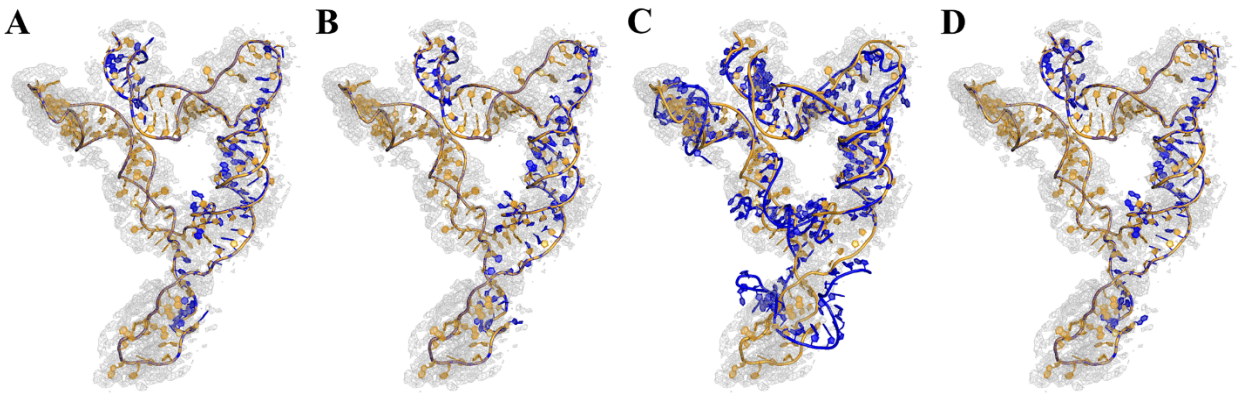



**Figure 3.10 Arena generates a structure from a semi-coarse-grained model that fits its experimental electron density.**
The native structure (orange) of the 146-nt 9S rRNA fragment (PDB 6SG9, chain CA) is overlaid with the resulting structure (blue) from Arena (A), PDBFixer (B), C2A (C), and rna_thread (D). The outputs were generated from the original semi-coarse-grained structure in the PDB. The electron density map (gray) is shown with contour level 3.0. Structures were visualized in PyMOL.

**Table 3.4 Performance of programs on reconstruction of 6SG9.**

| | Arena | PDBFixer | C2A | rna_thread | 6SG9 |
|---|---|---|---|---|---|
| **Q-score** | 0.34 ± 0.02 (0.59) | 0.33 ± 0.02 (0.03) | 0.04 ± 0.01 (8.68E-26) | 0.33 ± 0.02 (0.30) | 0.34 ± 0.02* |
| **RMSD (Å)** | 0.00 | 0.00 | 10.39 | 0.00 | - |
| **INF** | 0.76 | 0.88 | 0.77 | 0.29 | - |
| **Clashes** | 63 | 181 | 635 | 352 | - |
| **Runtime (s)** | 2.99 | 15.65 | 79.38 | 9.14 | - |
| **MAE (Å)** | 0.00 | 0.00 | 3.79 | 0.00 | - |

*p-value between the per-residue Q-scores for the original structure (6SG9) and the output generated by each respective program was calculated using a paired, two-tailed t-test.

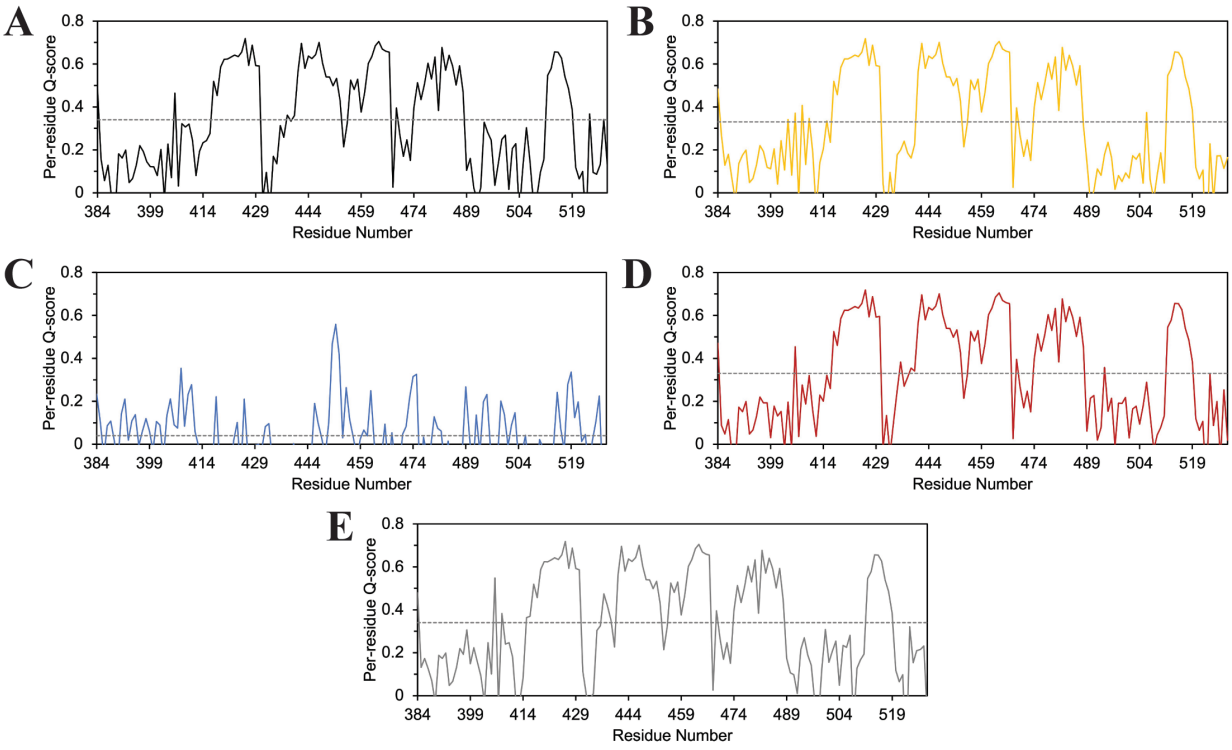



**Figure 3.11 Per-residue Q-scores for reconstruction of 6SG9.**
The per-residue Q-scores are shown for the structures generated by Arena (A), PDBFixer (B), C2A (C), rna_thread (D), and the original structure (E) compared to the experimental electron density. The average Q-score is shown as a dotted line. Q-scores were calculated using the MapQ plugin in Chimera.

## 3.5 Discussion

We developed the Arena program to build full atomic, refined structures of any coarse-grained RNA. On a benchmark of 361 experimental RNA structures, we found that Arena is fast (< 40 sec for all RNAs tested, with average runtimes < 3.5 sec), achieves low RMSDs (< 5.6 Å for all RNAs tested, with average RMSDs < 3.7 Å), has high INF (average 0.60-1), and reduces or completely removes all clashes compared to other programs (removes on average ≥ 77% of clashing atom pairs). Due to its fast runtime, Arena does not require a high-performance computing cluster and can be run locally. Arena is amenable to a range of RNA sizes, and we tested RNAs between 30 and 692 nt in the benchmarking dataset. Arena can be used to reconstruct low-

resolution experimental structures from cryo-EM, X-ray crystallography, and NMR or computationally predicted coarse-grained structures and is applicable for all types of RNAs.

A major reason for the high speed of Arena is that it performs superimposition followed by deterministic refinement based on ideal RNA geometries. This proves to be a more effective and accurate way to rebuild a full atomic structure compared to the other tested programs that perform stochastic simulations. At the same time, Arena has some limitations that could be addressed in future versions of the program. While the current runtime of Arena is sufficiently fast for most applications, the clash removal step is one of the slowest and could be improved by using a lattice coordinate system for the atoms [19]. To further increase the accuracy of the final structure, Arena could implement a check to fix improper ribose pucker conformations.

## 3.6 Materials and Methods

### Arena Algorithm

Arena first generates a full atomic structure from the coarse-grained input and then iteratively refines the placement of those atoms by correcting bond lengths and bond angles, base and base pair conformations, and pairwise atomic clashes. For the default Arena option, only the missing atoms added to the input structure are movable and fixed during the subsequent refinement steps. For the extended Arena option, the default Arena run is first performed; if the structure is discontinuous or atoms are still movable (because the performance metrics do not meet the specified thresholds), Arena performs an identical refinement but with all atoms allowed to move. Additional running options are described at https://github.com/pylelab/Arena#usage.

*Step 1: Generate full atomic structure via superimposition.* The program first fills in all missing atoms via superimposition of fragments from an ideal RNA helix onto the input atoms. The ideal RNA is from the A-form 3DNA fiber model [20] for RNA with twist 32.7º and rise 2.548 Å [13]. Fragments chosen from the ideal RNA match the nucleotide identities of the input, such that inconsistent base atoms are automatically excluded (for example, N6 for nucleotide U would be excluded). The fragments are superimposed onto the input target, so that the pre-existing atoms in the target are not moved. The superimposition algorithm uses three-dimensional least squares minimization to find the rotation matrix and translation vector that minimize the RMSD between the target and ideal RNA atoms, which is an extension of the Kabsch algorithm [21]. Arena performs four levels of superimposition: residue-level, base pair-level, base pair stack-level, and local fragment-level. To perform superimposition, there must be at least three pre-existing atoms from the input structure in the level of consideration. Ideally, we want to include the minimum number of residues that together have a total of three atoms. For example, if every residue in the input has three atoms, then only residue-level superimposition is necessary, in which the corresponding 1-nt fragment from the ideal RNA is superimposed onto each residue in the target. For residues with ≤ 2 atoms, the program proceeds to base pair-level filling based on the assigned secondary structure from CSSR [14]. There must be at least three atoms between the bases in the pair for superimposition to occur. If each residue in a pair only has one atom, the program checks a base pair stack, which includes the two residues adjacent to the current residue. If both adjacent residues are base-paired, the program uses a 3-nt helix (base-paired) fragment for the superimposition; if only one is base paired, the program used a 2-nt helix fragment corresponding to the base-paired residues. Finally, if the current residue is not base-paired and has ≤ 2 atoms, the

program uses local fragment-level superimposition based on the current residue and its two adjacent residues without considering base pairs. Residues without even a single atom are skipped.

*Step 2: Refine bond lengths and bond angles.* Arena checks every covalent bond connection (intra-residue and the inter-residue O3'-P bonds) and computes the bond length in 3D space. Arena implements a check to ensure residues are connected (in case of missing residues); sequentially adjacent residue pairs with at least one atom pair within 8 Å are considered connected. Arena calculates the fractional difference between the computed bond length and the ideal bond length taken from the Amber forcefield parameters in QRNAS [22]. If this is greater than the allowable 10% tolerance, then each atom is moved half of the difference between the calculated and ideal bond length. If one of the atoms is not movable because it was present in the input structure, then the movable atom is moved the full distance. Only bonds for which at least one atom is movable are checked. Refining bond angles happens by defining pseudobonds between the previous and subsequent atoms of the current atom. This distance is adjusted identically to bond lengths by comparing to an ideal pseudobond length $c$ calculated as $c = \sqrt{a^2 + b^2 - 2ab * cos(\theta)}$ , where $a$ and $b$ are the ideal bond lengths between the previous atom and subsequent atom to the current atom, respectively, and $\theta$ is the ideal bond angle taken from the Amber forcefield parameters in QRNAS [22]. Inter-nucleotide pseudobonds are also considered (C3'-P, O3'-OP1, O3'-OP2, and O3'-O5'). Arena increases or decreases the pseudobond length to indirectly adjust the angle.

*Step 3: Refine base and base pair conformations.* First, each base's conformation is checked individually by computing the RMSD between the coordinates of the current residue's base atoms (including the C1' atom) and an ideal 1-nt fragment from the ideal RNA. To compute the RMSD,

the ideal base is superimposed onto the target by a least squares fit [21]. If the RMSD ≥ 0.1 Å, then the coordinates of the target are moved to match those from the superimposed ideal base. Second, each Watson-Crick base pair assigned from CSSR [14] is checked (A:U and G:C). The ideal base pair from the ideal RNA helix is superimposed onto the target, and the RMSD is calculated in the same way. If the RMSD ≥ 0.28 Å, the coordinates of the target base pair are moved to match those from the superimposed ideal base pair. The RMSD thresholds were chosen from those used in DSSR [23]. To reduce running time, the CSSR secondary structure assignment, which is required in base pair conformation correction, is updated every tenth iteration.

*Step 4: Remove clashing atom pairs.* Arena first iterates through every possible residue combination, calculating the 3D distance between the C1' atoms. If the C1'-C1' distance is > 18.16 Å for the nucleotide pair, it is not physically possible for the constituent atoms to clash (**Text S3**). If this condition is not met, Arena proceeds to iterate through all inter-residue atom pairs and calculates their 3D distance. The distance threshold that defines a clash is calculated as the sum of the van der Waals radii of the two atoms minus 0.4 Å [24]. If the atom-atom distance is less than the threshold, they are considered to sterically clash, and the atoms are moved apart as described in *Step 2*.

The program iterates through *Steps 2-4* to generate a refined structure. The number of iterations increases with the length of the RNA according to a linear function, where the maximum number of iterations is $0.4*L + 100$, where $L$ is the length of the RNA in nucleotides. The program finishes after reaching the maximum number of iterations or until no more atoms are moved, whichever comes first.

**Running NAST and SimRNA**

To perform structure prediction by NAST, 20,000 steps of NAST simulation were run at 300 Kelvin, using the random seed 11111111 and secondary structure prediction from PETfold. This generates C3'-only coarse-grained RNA structures, which are then reconstructed into full atomic models using either NAST's built-in module (C2A) or external programs (Arena and PDBFixer). Only 356 out of the 361 RNAs were used for full atomic reconstruction due to an indexing error with C2A.

To perform structure prediction by SimRNA (version 3.20), the following parameters recommended by the SimRNA user manual were used to perform the Simulated Annealing Monte Carlo simulation guided by the PETfold-predicted secondary structure:

```
NUMBER_OF_ITERATIONS          16000000
TRA_WRITE_IN_EVERY_N_ITERATIONS 16000

INIT_TEMP 1.35
FINAL_TEMP 0.90

BONDS_WEIGHT 1.0
ANGLES_WEIGHT 1.0
TORS_ANGLES_WEIGHT 0.0
ETA_THETA_WEIGHT 0.40
```

Conformations generated by the SimRNA simulation were clustered at an RMSD threshold of 5.00 Å to obtain the representative structure of the largest cluster. This structure only contains the following atoms: P, C4', N9, C2 and C6 for nucleotides G and A; P, C4' N1, C2 and C4 for nucleotides C1 and U. This coarse-grained structure is then reconstructed into a full atomic model using either SimRNA's internal module (SimRNA_trafl2pdbs) or external programs (Arena and PDBFixer). Only 352 out of 361 RNAs can complete the SimRNA run due to a limitation on maximum running time (7 days) on our supercomputer cluster.

Before calculating the RMSD between the model and native structure, the TM-score program is used to superimpose the model to native.

**Calculation of Interaction Network Fidelity**

To calculate the Interaction Network Fidelity (INF) between the output and each native structure, we used the INF module in the RNA_assessment package provided by the RNA-Puzzles challenge (https://github.com/RNA-Puzzles/RNA_assessment). We modified the script to take both the native structure and output structure generated by each program. We first added chain IDs to each structure and reindexed them to be in the correct format.

```
$ python3 RNA_assessment/INF.py RNA_assessment/example/target.native.pdb target.index
RNA_assessment/example/target.output.pdb target.index
```

In structures where no base-base interactions are found, we arbitrarily assign an INF score of 0.

**Calculation of Clash Cutoff**

In *Stage 4*, iterating through all possible atom pairs is computationally expensive. Thus, Arena increases speed by implementing a check that calculates if it is physically possible for any atoms in a particular residue pair to clash. Arena calculates the C1'-C1' distance for a residue pair, and if the distance exceeds a threshold of 18.161, then a clash is not possible between any atoms in those residues. Thus, Arena will proceed to the next residue pair and will not iterate between all the atom pairs in those residues. If the distance is ≤ 18.161, then it is possible for a clash to occur, so Arena proceeds to check all inter-residue atom pairs. The radius of a residue represented as a sphere can be found by adding the pseudobond distances between nonconsecutive backbone atoms to the van der Waals radius of the largest atom, which is the P atom. The threshold is simply twice this radius:

$$2 * (< C1' - C4' > + < C4' - O5' > + < O5' - OP1 > + < \text{VDW radius of P} >)$$

$$2 * (2.343 + 2.406 + 2.532 + 1.8) = 18.16$$

## 3.7 SUPPLEMENTARY DATA

**Table 3.5 Performance of programs on backbone reconstruction from P, C1', and base atoms.**

| | Arena | PDBFixer | C2A | RCrane |
|---|---|---|---|---|
| **RMSD (Å)** | 0.49 ± 0.004* | 0.97 ± 0.02 (9.32E-107) | 4.77 ± 0.16 (3.83E-88) | 0.38 ± 0.005 (2.06E-105) |
| **INF** | 1.00 ± 0.00* | 1.00 ± 0.00 (1) | 0.59 ± 0.00 (3.03E-242) | 1.00 ± 0.00 (0.32) |
| **Clashes** | 17.53 ± 1.56* | 109.46 ± 6.58 (1.98E-45) | 667.51 ± 38.91 (7.66E-48) | 77.18 ± 4.83 (6.91E-44) |
| **Runtime (s)** | 2.75 ± 0.24* | 511.75 ± 92.83 (7.38E-8) | 169.43 ± 45.19 (2.49E-4) | 193.52 ± 9.68 (4.62E-61) |
| **MAE (Å)** | 0.12 ± 0.00* | 0.25 ± 0.00 (6.87E-229) | 1.75 ± 0.06 (8.62E-91) | 0.09 ± 0.00 (1.02E-105) |

Average values are reported with standard error of the mean. *p-value between Arena and each program was calculated using a paired, two-tailed t-test.

**Table 3.6 Performance of programs on reconstruction from C3' atoms.**

| | Arena | PDBFixer | C2A |
|---|---|---|---|
| **RMSD (Å)** | 2.98 ± 0.02* | 5.64 ± 0.02 (1.92E-258) | 5.08 ± 0.14 (2.30E-41) |
| **INF** | 0.69 ± 0.00* | 0.01 ± 0.00 (0) | 0.54 ± 0.00 (7.50E-105) |
| **Clashes** | 0.23 ± 0.03* | 275.67 ± 14.77 (7.43E-55) | 606.22 ± 41.33 (1.65E-38) |
| **Runtime (s)** | 2.93 ± 0.24* | 979.06 ± 133.79 (1.76E-12) | 85.65 ± 18.93 (1.36E-5) |
| **MAE (Å)** | 0.88 ± 0.01* | 2.24 ± 0.00 (1.42E-298) | 1.81 ± 0.06 (5.68E-49) |

Average values are reported with standard error of the mean. *p-value between Arena and each program was calculated using a paired, two-tailed t-test.

**Table 3.7 Performance of programs on reconstruction from glycosidic N atoms.**

| | Arena | PDBFixer |
|---|---|---|
| **RMSD (Å)** | 2.74 ± 0.02* | 4.56 ± 0.02<br>(2.08E-205) |
| **INF** | 0.74 ± 0.00* | 0.04 ± 0.00<br>(0) |
| **Clashes** | 2.24 ± 0.17* | 528.67 ± 26.19<br>(4.69E-61) |
| **Runtime (s)** | 3.48 ± 0.23* | 958.64 ± 130.03<br>(1.28E-12) |
| **MAE (Å)** | 0.94 ± 0.01* | 1.92 ± 0.01<br>(5.98E-250) |

Average values are reported with standard error of the mean. *p-value between Arena and each program was calculated using a paired, two-tailed t-test.

**Table 3.8 Performance of programs on reconstruction from NAST coarse-grained predictions.**

| | Arena (default) | Arena (extended) | PDBFixer | C2A |
|---|---|---|---|---|
| **RMSD (Å)** | 39.45 ± 1.60* | 39.45 ± 1.60<br>(8.31E-5) | 40.01 ± 1.59<br>(8.94E-44) | 39.73 ± 1.60<br>(2.82E-7) |
| **INF** | 0.48 ± 0.01* | 0.48 ± 0.01<br>(0.14) | 0.01 ± 0.00<br>(2.28E-234) | 0.39 ± 0.00<br>(1.85E-51) |
| **Clashes** | 2.40 ± 0.25* | 0.18 ± 0.03<br>(8.00E-18) | 270.38 ± 12.91<br>(2.00E-64) | 646.25 ± 40.41<br>(9.25E-44) |
| **Runtime (s)** | 3.11 ± 0.19* | 3.99 ± 0.38<br>(2.81E-5) | 882.90 ± 117.32<br>(4.81E-13) | 103.12 ± 17.60<br>(2.19E-8) |
| **MAE (Å)** | 15.72 ± 0.64* | 15.72 ± 0.64<br>(1.67E-5) | 16.22 ± 0.63<br>(2.67E-104) | 16.06 ± 0.63<br>(1.83E-25) |

Average values are reported with standard error of the mean. *p-value between Arena (default) and each program (or the Arena extended option) was calculated using a paired, two-tailed t-test.

**Table 3.9 Performance of programs on reconstruction from SimRNA coarse-grained predictions.**

| | Arena (default) | Arena (extended) | PDBFixer | SimRNA |
|---|---|---|---|---|
| **RMSD (Å)** | 32.92 ± 1.11* | 32.93 ± 1.11 (1.66E-26) | 32.92 ± 1.11 (0.54) | 32.94 ± 1.11 (0.25) |
| **INF** | 0.60 ± 0.01* | 0.61 ± 0.01 (1.00E-2) | 0.43 ± 0.00 (6.80E-132) | 0.60 ± 0.01 (0.62) |
| **Clashes** | 198.85 ± 7.79* | 0.14 ± 0.02 (5.31E-82) | 487.25 ± 17.96 (6.51E-90) | 411.88 ± 15.79 (7.31E-84) |
| **Runtime (s)** | 2.92 ± 0.13* | 3.14 ± 0.17 (1.40E-4) | 452.93 ± 56.19 (1.53E-14) | - |
| **MAE (Å)** | 12.97 ± 0.48* | 12.92 ± 0.48 (0.32) | 12.99 ± 0.48 (6.80E-40) | 12.92 ± 0.48 (0.35) |

Average values are reported with standard error of the mean. The run time for SimRNA's built-in full-atomic reconstruction program is not reported, as we could not isolate this submodule of SimRNA from the rest of the SimRNA pipeline. *p-value between Arena (default) and each program (or the Arena extended option) was calculated using a paired, two-tailed t-test.

**Table 3.10 Performance of programs on reconstruction of 6Q9A.**

| | Arena (default) | Arena (extended) | PDBFixer | SimRNA |
|---|---|---|---|---|
| **RMSD (Å)** | 110.59 | 110.60 | 110.57 | 110.45 |
| **INF** | 0.37 | 0.38 | 0.31 | 0.37 |
| **Clashes** | 814 | 0 | 1788 | 1731 |
| **Runtime (s)** | 10.63 | 17.22 | 6001.07 | - |
| **MAE (Å)** | 45.05 | 45.05 | 45.05 | 45.02 |

**Table 3.11 Performance of Arena on NMR models.**

| | 2N3Q | 2KOC |
|---|---|---|
| **Length (res)** | 61 | 14 |
| **Models** | 21 | 20 |
| **RMSD (Å)** | 3.68 ± 0.03 | 3.32 ± 0.03 |
| **INF** | 0.45 ± 0.02 | 0.52 ± 0.03 |

| | | |
|---|---|---|
| **Clashes** | 0.05 ± 0.05 | 0.10 ± 0.07 |
| **Runtime (s)** | 0.75 ± 0.06 | 0.14 ± 0.01 |
| **MAE (Å)** | 1.22 ± 0.01 | 1.17 ± 0.02 |

Average values are reported with standard error of the mean.

**Table 3.12 Comparison of Arena's performance on different classes of RNA.**

| | **tRNA** | **rRNA** | **snRNA** | **ribozyme and self-splicing intron** | **riboswitch and aptamer** | **other** | **all** |
|---|---|---|---|---|---|---|---|
| **Structures** | 88 | 46 | 20 | 31 | 70 | 106 | 361 |
| **RMSD (Å)** | 3.72 ± 0.04 (0.09) | 3.78 ± 0.08 (0.05) | 4.01 ± 0.09 (0.00) | 3.43 ± 0.07 (0.03) | 3.48 ± 0.06 (0.01) | 3.59 ± 0.07 (0.34) | 3.63 ± 0.03 |
| **INF** | 0.60 ± 0.01 (0.94) | 0.55 ± 0.01 (1.24E-5) | 0.62 ± 0.02 (0.32) | 0.64 ± 0.01 (0.02) | 0.61 ± 0.01 (0.31) | 0.61 ± 0.01 (0.67) | 0.60 ± 0.01 |
| **Clashes** | 0.18 ± 0.04 (0.36) | 0.67 ± 0.36 (0.01) | 0.35 ± 0.30 (0.78) | 0.58 ± 0.30 (0.13) | 0.19 ± 0.08 (0.45) | 0.15 ± 0.05 (0.17) | 0.28 ± 0.06 |
| **MAE (Å)** | 1.15 ± 0.02 (0.29) | 1.21 ± 0.03 (0.01) | 1.25 ± 0.04 (0.01) | 1.05 ± 0.02 (0.03) | 1.06 ± 0.02 (0.00) | 1.13 ± 0.03 (0.80) | 1.13 ± 0.01 |

Average values are reported with standard error of the mean. All metrics are from the results of the reconstruction from P atoms by Arena. *p-value was calculated between each subset and the remaining dataset using an unpaired (equal variance), two-tailed t-test.

## 3.8 References


1. Ma, H., Jia, X., Zhang, K. & Su, Z. (2022). Cryo-EM advances in RNA structure determination. *Signal Transduct Target Ther*. **7**, 58.

2. Pyle, A. M. (2016). Group II Intron Self-Splicing. *Annu Rev Biophys*. **45**, 183-205.

3. Childs-Disney, J. L., Yang, X., Gibaut, Q. M. R., Tong, Y., Batey, R. T. & Disney, M. D. (2022). Targeting RNA structures with small molecules. *Nat Rev Drug Discov*. **21**, 736-762.

4. Pyle, A. M. & Schlick, T. (2016). Challenges in RNA Structural Modeling and Design. *J Mol Biol*. **428**, 733-735.

5. Keating, K. S. & Pyle, A. M. (2010). Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proc Natl Acad Sci U S A*. **107**, 8177-8182.

6. Jonikas, M. A., Radmer, R. J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. & Altman, R. B. (2009). Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*. **15**, 189-199.

7. Boniecki, M. J., Lach, G., Dawson, W. K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K. M. & Bujnicki, J. M. (2016). SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res*. **44**, e63.

8. Rotkiewicz, P. & Skolnick, J. (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem*. **29**, 1460-1465.

9. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R. & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. **13**, e1005659.

10. Jonikas, M. A., Radmer, R. J. & Altman, R. B. (2009). Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics*. **25**, 3259-3266.

11. Cheng, C. Y., Chou, F. C. & Das, R. (2015). Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol*. **553**, 35-64.

12. Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., Berman, H. M. & Consortium, R. N. A. O. (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*. **14**, 465-481.

13. Chandrasekaran, R. & Arnott, S. (1989). The structures of DNA and RNA helices in oriented fibers. *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, Group VII/1b, Nucleic Acids*. 31-170.

14. Zhang, C. & Pyle, A. M. (2022). CSSR: assignment of secondary structure to coarse-grained RNA tertiary structures. *Acta Crystallogr D Struct Biol*. **78**, 466-471.

15. Zhang, C., Zhang, Y. & Pyle, A. M. (2022). rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling. *Journal of Molecular Biology*. 167904.

16. Cruz, J. A., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., Cao, S., Das, R., Ding, F., Dokholyan, N. V., Flores, S. C., Huang, L., Lavender, C. A., Lisi, V., Major, F., Mikolajczak, K., Patel, D. J., Philips, A., Puton, T., Santalucia, J., Sijenyi, F., Hermann, T., Rother, K., Rother, M., Serganov, A., Skorupski, M., Soltysinski, T., Sripakdeevong, P., Tuszynska, I., Weeks, K. M., Waldsich, C., Wildauer, M., Leontis, N. B. & Westhof, E. (2012). RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*. **18**, 610-625.

17. Petersen, N. P. (2020). Computational methods for RNA 3D structure prediction and comparison. Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

18. Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M. F. & Chiu, W. (2020). Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods*. **17**, 328-334.

19. Zhang, Y., Kolinski, A. & Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J*. **85**, 1145-1164.

20. Li, S., Olson, W. K. & Lu, X.-J. (2019). Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Research*. **47**, W26-W34.

21. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*. **32**, 922-923.

22. Stasiewicz, J., Mukherjee, S., Nithin, C. & Bujnicki, J. M. (2019). QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct Biol*. **19**, 5.

23. Lu, X.-J., Bussemaker, H. J. & Olson, W. K. (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*. **43**, e142-e142.

24. Chen, V. B., Arendall, W. B., III, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*. **66**, 12-21.

# 4. CHAPTER 4: CONCLUSIONS

## 4.1 ADDITIONAL CONTRIBUTIONS

In addition to the work presented in this dissertation, I had the opportunity to contribute to several collaborative projects related to RNA structure. These studies are listed below with citations and a brief description of my role.

Xiao, W., Hann, T., **Perry, Z.**, Pandey, V., Cheng, J., Wohlschlegel, J., Pyle, A. M., Black, D. (2026). The long noncoding RNA Malat1 contains an internal ribosome entry site mediating micropeptide translation. Submitted to *Science Advances*.

**Contribution:** Assisted with data analysis and interpretation of the MALAT1 secondary structure and edited the corresponding manuscript sections.

Mahadeshwar, G., Tavares, R. C. A., Wan, H., **Perry, Z. R.**, & Pyle, A. M. (2023). RSCanner: rapid assessment and visualization of RNA structure content. *Bioinformatics (Oxford, England)*, 39(3), btad111.

**Contribution:** Assisted with testing RSCanner parameters on different RNAs, figure preparation, and manuscript editing.

Kumar, A., Wan, H., **Perry, Z.**, Patel, S., Tavares, R., Pyle, A. M. (2023). Discovery of a Well-Folded Protein Interaction Hub Within the Human Long Non-Coding RNA NORAD. *bioRxiv*.

**Contribution:** Assisted with protein purification and binding assays for the NORAD-Pumilio1-FUBP3 complex.

## 4.2 Summary of Results

This dissertation investigated multiple levels of noncoding RNA (ncRNA) structure using experimental and computational approaches. **Chapter 2** presented the secondary structure of a long noncoding RNA (lncRNA) in breast cancer cells, and **Chapter 3** presented a full-atomic, three-dimensional reconstruction program that can be applied to diverse ncRNAs.

### *4.2.1 HOTAIR*

The lncRNA HOTAIR is involved in promoting breast cancer progression, but little is known about its mechanism. Since RNA structure influences mechanism, its secondary structure was determined in both a metastatic breast cancer cell line and *in vitro*. This involved implementing key experimental approaches, including the chemical probing technique SHAPE-MaP with the reagent 2A3 and a breast cancer line engineered to overexpress HOTAIR. Chemical probing reactivities were used to constrain secondary structure predictions, revealing that HOTAIR folds into four major structural domains in both cellular and *in vitro* contexts. The first two domains exhibit high similarity in both contexts, while subtle differences observed in Domains 3 and 4 show that the cellular context influences RNA architecture.

Beyond the overall architecture, local structural features were identified by applying SHAPE reactivity and Shannon entropy criteria to identify the most well-structured, rigid, and/or conformationally heterogeneous regions. Some of these regions were unique to the cellular context and represent candidate functional structures. Direct comparison between *in cellulo* and *in vitro* probing profiles identified clusters of nucleotides that are either more protected or accessible in cells, which may be due to intermolecular interactions or structural rearrangements specific to the cellular context.

Finally, conservation analysis across unannotated primate genomes identified putative HOTAIR homologs in 190 species. Multiple sequence alignments in conjunction with the *in cellulo* structure prediction were used to predict base pairs with statistically significant covariation, indicating the presence of conserved structural elements in HOTAIR.

This work demonstrates that the cellular environment influences RNA structure, especially in a cancer context, where specific structures may be driving the disease phenotype. For HOTAIR, its structure may be involved in regulating protein binders that control gene expression or epigenetic silencing. The conservation analysis further shows that HOTAIR is highly conserved in primates. There are many other lncRNAs for which this structure-first approach would be a valuable starting point for understanding their functional or disease relevance.

### *4.2.2 Arena*

Experimentally determined three-dimensional (3D) RNA structures often lack full-atomic resolution, limiting mechanistic interpretation. To reconstruct full-atomic RNA structures from low-resolution, coarse-grained models, the open-source software Arena was developed. It allows the user to build an accurate model within seconds starting from as little as one atom per nucleotide. Arena first reconstructs missing non-hydrogen atoms using a model RNA structure and iteratively refines their placement by correcting bond lengths, bond angles, base conformations, base pair conformations, and steric clashes. Optimization tests showed that scaling the maximum number of iterations with the length of the RNA provided the best balance between computational runtime and accuracy.

This program was benchmarked on a large dataset of diverse ncRNAs, including tRNAs, rRNAs, snRNAs, and riboswitches. Several benchmarking datasets were generated from different

coarse-grained representations, including structures containing only phosphorus atoms, only glycosidic nitrogen atoms, or only backbone atoms, etc. Across the tested inputs, Arena outperformed other reconstruction programs with higher accuracy, reduced clashes, and faster run time. Some alternative programs perform well in specific use cases, such as Rosetta's rna_thread, but it requires structures containing all backbone atoms. The advantage of Arena is that only one atom is required per nucleotide, regardless of its identity. Case studies on two rRNA structures further validated the utility of Arena in reconstructing accurate models.

This work presents a fast and highly accurate program for full-atomic reconstruction. Generating more accurate 3D structures will enable greater mechanistic insight, since structural models can be used to guide experimental studies or for all-atom computational simulations. Making tools like Arena accessible to the RNA community will hopefully support continued structure-function investigations of ncRNAs.

## 4.3 Future Directions and Perspectives

### *4.3.1 Towards a Mechanism of HOTAIR*

To better understand diseases like cancer, it is necessary to move beyond a strictly protein-centric point of view to consider the role of RNA in disease regulatory networks. Although HOTAIR is only one facet of cancer progression, it is an important example of RNA-based gene regulation. There is also evidence for HOTAIR forming alternative isoforms, some of which may have distinct structures and functions (See **Figure 1.2**) [1]. To test the structure-function relationship in HOTAIR or any of its isoforms, robust functional assays are needed, such as a tumor spheroid invasion assay [2, 3]. One could target specific structural elements with antisense structure-disrupting nucleotides and measure the resulting effect on invasion. Furthermore,

characterizing HOTAIR isoforms and their structures in patient-derived xenograft cell lines [4] or primary patient cells with high HOTAIR expression [5] would help connect biochemical mechanism with clinical relevance. Performing *in vivo* SHAPE-MaP directly on xenograft mouse tumors, where human HOTAIR is expressed, would be a clinically meaningful but technically challenging frontier.

With an *in cellulo* structure now available, future work can directly test protein interactions with specific structural elements to determine mechanism. This can be done through protein discovery approaches, such as RNA structure pull downs followed by mass spectrometry, after which hits can be validated through biochemical binding assays [6]. Ultimately, determining the structure of a protein in complex with a HOTAIR structural module using cryo-EM would provide the greatest mechanistic insight. Further work can confirm the presence of novel modifications on HOTAIR and explore how they modulate structure or influence protein binding. Although our mechanistic understanding of HOTAIR is incomplete, it is just one of many lncRNAs awaiting rigorous characterization, and structure determination and analysis provide a key foundation for such future studies.

### *4.3.2 Secondary Structure Methods*

Chemical probing approaches such as SHAPE-MaP are a fast and powerful way to determine the structure of any reasonably abundant RNA in living cells. However, one should approach secondary structure predictions with caution, given they are a static model, whereas RNAs are part of dynamic conformational ensembles. Probing also reports on the average structural state across time and space and does not give information about dynamical structure changes occurring under different conditions that may regulate function. This could partially be

addressed biochemically by applying chemical probing to subcellular fractions or by comparing probing data at different time points. Some computational approaches have also been developed to deconvolute distinct structural states in averaged probing data [7].

The criteria used in this work to identify structural modules were SHAPE reactivity and Shannon entropy. While these linear metrics are a useful initial assessment, they likely overlook important structural features. An alternative approach that would require benchmarking would be to convert an RNA secondary structure prediction into a graph representation [8]. These could be analyzed in many ways, such as by calculating radius of gyration or fractal dimension, which could reveal structural patterns and motifs across different classes of long RNAs. It may also prove useful for parsing through several kilobase-long viral RNA genomes, which often contain high structural density relative to mRNA [9]. However, these analyses ultimately depend on the input quality of the structure prediction. Thus, further improvements are needed in prediction algorithms for RNA structure, which could involve designing algorithms that combine biophysical and thermodynamic first principles with machine learning approaches.

### *4.3.3 Computational Predictions*

As discussed in *Section 4.3.2*, improved prediction algorithms are needed for RNA structure, particularly at the tertiary structure level, since RNA structure prediction lags behind protein prediction. Although the program Arena can successfully reconstruct and refine structures from initial atomic coordinates, it cannot build nucleotides *de novo*. Many approaches for RNA structure prediction have been developed, such as those relying on physics principles, homology modeling, deep learning, or a combination thereof, none of which fully captures the complexity of RNA folding alone [10]. Arena, which does not rely on stochastic sampling or machine

learning, demonstrates that simple geometric constraints can lead to highly accurate structures. At the same time, deep learning approaches can predict *de novo* structures with impressive accuracy. Arena's capabilities could further be improved by integrating deep learning-based folding predictions before or after Arena's refinement algorithm. In general, combining the strengths of geometric constraint-based and deep learning approaches may help address the fundamental problem of understanding RNA folding principles. However, structure alone is uninformative until it is connected to biological function, so predicting function directly from structure is also a high-priority computational challenge.

### *4.3.4 Final Perspectives*

Regulatory RNA structures may promote cancer and other diseases, making them promising candidates for structure-targeting therapies. As in the protein engineering field, novel RNAs could also be designed with custom structures or altered protein binding specificities for industrial and biomedical applications. Continuing to develop tools to study, determine, and interpret RNA structure will help unveil the mysteries of many ncRNAs and their fundamental patterns of folding and function. At the same time, RNA is only one component of highly interconnected cellular networks. Although reductionist biochemical approaches are important for determining individual mechanisms, it is the intricate interactions of various molecular components that lead to their emergent functions. As molecular biology discoveries continue to illuminate the complexity of macromolecules, may we never lose our *sense of wonder* for the natural world.

## 4.4 References


1. Potolitsyna, E., Hazell Pickering, S., Tooming-Klunderud, A., Collas, P. & Briand, N. (2022). De novo annotation of lncRNA HOTAIR transcripts by long-read RNA capture-seq reveals a differentiation-driven isoform switch. *BMC Genomics*. **23**, 658.

2. Cavaco, M., Fraga, P., Valle, J., Andreu, D., Castanho, M. & Neves, V. (2021). Development of Breast Cancer Spheroids to Evaluate Cytotoxic Response to an Anticancer Peptide. *Pharmaceutics*. **13**.

3. Vinci, M., Box, C. & Eccles, S. A. (2015). Three-dimensional (3D) tumor spheroid invasion assay. *J Vis Exp*. e52686.

4. Finlay-Schultz, J., Jacobsen, B. M., Riley, D., Paul, K. V., Turner, S., Ferreira-Gonzalez, A., Harrell, J. C., Kabos, P. & Sartorius, C. A. (2020). New generation breast cancer cell lines developed from patient-derived xenografts. *Breast Cancer Res*. **22**, 68.

5. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S. & Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. **464**, 1071-1076.

6. Kumar, A., Wan, H., Perry, Z., Patel, S., Tavares, R. & Pyle, A. M. (2023). Discovery of a Well-Folded Protein Interaction Hub Within the Human Long Non-Coding RNA NORAD. *BioRxiv*.

7. Aviran, S. & Incarnato, D. (2022). Computational Approaches for RNA Structure Ensemble Deconvolution from Structure Probing Data. *J Mol Biol*. **434**, 167635.

8. Schlick, T. (2018). Adventures with RNA graphs. *Methods*. **143**, 16-33.

9. Tavares, R. C. A., Mahadeshwar, G., Wan, H., Huston, N. C. & Pyle, A. M. (2021). The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. *J Virol*. **95**.

10. Kagaya, Y., Liu, B. & Kihara, D. (2026). Computational approaches for RNA structure prediction and design. *Cell Reports Physical Science*.