Second Year Progress Report
Grace Tully

## I. Introduction

Recent market reports expect the demand for wearable health devices to soar over the next five years to anywhere between 150 to 300 billion due to increased awareness and desire for autonomy over personal health decisions[1]. Consequently, there is continued effort in advancing nanotechnology for developing wearable devices that can monitor fluctuations in trace concentrations of certain biomarkers. Aptamer-based sensors that model Nature's use of complex RNA structures for feedback regulation of metabolic processes have been proposed as a solution to this unmet need [2]. However, advances in the underlying technology of aptamer devices have outpaced the discovery of high affinity aptamers themselves, limiting their level of accuracy and ultimately commercialization for clinical diagnostics [3].

The current and only method for aptamer selection over the past 30 years is SELEX (Selective evolution of ligands by exponential enrichment) [4]. Due to the small molecular weight and chemical similarity of small molecule compounds, the most general SELEX protocol has had limited success discovering high affinity aptamers ($K_D \sim$ nM) for small molecules, prompting decades of work toward SELEX modifications to improve outcomes [5]. Even once a high affinity aptamer has been discovered, it often lacks the stability and specificity necessary to be incorporated in a wearable device to accurately distinguish between chemically similar metabolites in a sample of bodily fluids.

Over the past ten years, growth in precision in RNA sequencing technology has transformed our understanding of chemical biology. Concurrent advances in high-throughput RNA structure have recently been used to successfully screen for RNA-targeting drug candidates [6]. The Das Lab has recognized the power of these high-throughput protocols to collect millions of RNA structural data to train a foundation model, RibonanzaNet, that predicts RNA secondary structures and tertiary structure motifs [7]. However, no one has leveraged deep learning to test whether RNA chemical probing data can reveal binding affinities or predict sequences that preferentially bind small molecules.

This work proposes a model trained on chemical probing data that encodes small molecule inputs and predicts RNA aptamer sequences with the highest levels of affinity. If successful, the final model will overcome the limitations of traditional SELEX approaches and enhance the discovery of selective and specific aptamers for small molecule metabolites.

## II. Methods

### Experimental Design

*RNA Library*

Decades of oligonucleotide library design for initial SELEX pools found that increasing structural complexity increases the probability of finding selective aptamers. Initial libraries of oligonucleotide pools used for SELEX typically include $10^{16}$ unique sequences. Despite the large sequence space, statistical analyses have demonstrated that complex structural motifs in these starting libraries generated from parallel random synthesis are underrepresented. To circumvent this problem, it is common among SELEX practitioners to add a constant region of structural complexity to every sequence in the library to increase frequency of complex tertiary interactions, and consequently, binding of more specific aptamers [8]. However, for my proposed chemical mapping aptamer-search protocol, I cannot apply the same techniques for library design because of the constraint on sequencing capacity. Even with a high-throughput sequencer generating up to 20 billion reads, achieving high coverage (10,000 reads per unique modified sequence, barcoded by ligand and concentration condition) limits the experiment to a maximum of 2 million unique sequences per run. For my initial experiment and data collection, I will chemically probe the entire library with 32 distinct molecules, each evaluated across 6 different concentration conditions, which even further limits the unique sequence space to around 10,000 designs. Although this represents only a fraction of the sequences found in an initial randomized oligonucleotide SELEX pool, I believe I can match or exceed the level of structural complexity by using the custom-designed OpenKnot-7 100mer library. This library consists of 12,000 unique sequences, combining both naturally occurring RNA sequences and 8,000 synthetic designs from Eterna, a crowd-sourced database where human players engineer sequences to align with thermodynamic parameters favoring intricately folded motifs. Furthermore, this library's RNA length of 100 nucleotides is optimal for aptamer design by balancing structural complexity with folding stability. While longer sequences enhance the likelihood of forming motifs like kissing loops, pseudoknots, and three-way junctions, overly extended sequences also increase the risk of misfolding into kinetically trapped states.

*Small molecule library*

Ultimately this model will be trained on data reflecting the binding affinities of *in vitro* RNA structures to detect small molecules. The greatest utility of these aptamers would be incorporation in *ex vivo* devices that can detect the concentration of small molecules from a sample of blood or sweat. This makes aptamer selectivity especially challenging considering the chemical similarity of many small molecule metabolites that trigger different biochemical processes. Unlike the RNA library where structure and sequence diversity was maximized, I will initially choose 32 molecules from Selleckchem's Human Endogeneous Metabolite Compound Library to enhance the model's sensitivity for detecting even small chemical differences between metabolites found in human sweat and blood samples.

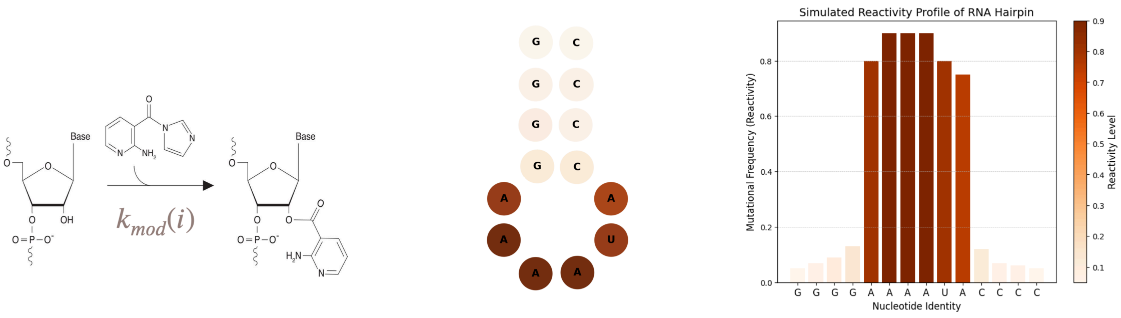

**Figure 2:** i) The electrophilic SHAPE reagent 2A3 (2-aminopyridine-3-carboxylic acid imidazolide) acylating the 2' hydroxyl (adapted from Marinus et al.) of a nucleotide. ii) A simple secondary structure of a stem-loop motif and a simulated reactivity profile.

## Chemical Probing Protocol

I plan to use a high throughput SHAPE-MaP (selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling) protocol to collect RNA structural probing data [9]. This method leverages the linear relationship between the nucleophilic 2' hydroxyl's chemical reactivity and the nucleotide's likelihood of being base-paired or shielded within intricate tertiary structures. Upon treatment with an electrophilic reagent, nucleotides will be acylated with a frequency that correlates with backbone flexibility. A nucleotide's acylation rate can be detected by the mutational frequency of its reverse transcribed base pair in complementary DNA strands sequenced after reverse transcription. The mutational frequencies for each nucleotide in the sequence are then converted to SHAPE "reactivity" vectors used to determine the RNA's secondary structure **(Figure 2)** .

Multiple groups have used differences in SHAPE reactivities (ΔSHAPE) between RNAs treated with and without molecules to detect small-molecules that bind specifically to RNA motifs. None have yet extended this work to correlate the magnitude of ΔSHAPE with the binding affinity of an RNA to a small molecule. However, an indirect measurement should be straightforward considering an RNA's reactivity profile reflects the average structure of the RNA ensemble. Under a constant concentration of an individual RNA in each sample, the average ΔSHAPE of impacted nucleotides should increase as the ligand is titrated until every RNA in the ensemble has bound the ligand. Once the average ΔSHAPE stabilizes as the ligand concentration increases (ΔΔSHAPE = 0), it can be assumed that the average reactivity of the ensemble reflects that of the ligand-bound RNA structure. Assuming a hyperbolic dependence of the average ΔSHAPE on ligand concentration, the aptamer's $K_D$ can be estimated by fitting the data to a hyperbolic binding curve. To confirm this prediction, I will perform a preliminary chemical probing experiment with an aptamer of known binding affinity to a ligand under different ligand concentrations, and use the ΔSHAPE reactivities to estimate the fraction of bound aptamers and the corresponding $K_{D.}$ This experiment would provide a valuable result for the scientific community as it could reveal a novel method for determining binding affinities of RNAs to ligands.

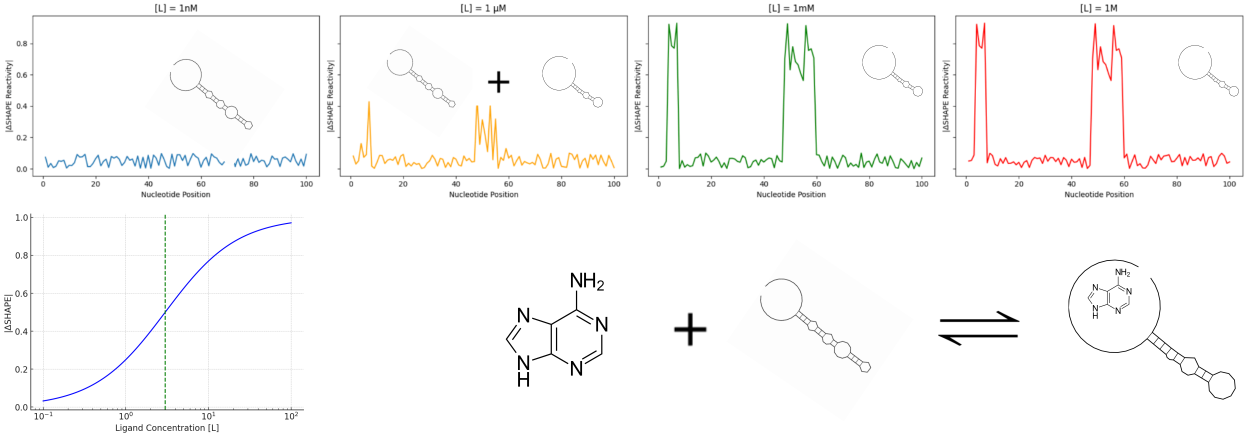

**Figure 3:** Simulated data showing theoretical prediction of how an aptamer's binding affinity can be determined from fitting the ΔSHAPE reactivities at different ligand concentrations to a hyperbolic curve.

Unfortunately for this large scale data collection, I cannot apply this simple thermodynamic model used to estimate binding affinities because all RNA aptamer candidates compete to bind ligands in a single sample, lowering the effective ligand concentration. Thus, primary consideration is optimizing the concentration of RNA and small molecule metabolites in each sample to ensure that binding affinities can be estimated based on initial sample conditions and corresponding reactivity profiles. In this protocol, I will optimize PCR amplification post reverse transcription of the designed DNA oligonucleotide library to generate at least 400 µL of RNA at a concentration of 120 ng/µL. Given that the average molecular weight of a 177-nt RNA (including barcodes and primers) is 58.6 kDa, the final total RNA molar concentration in each sample will be 2.25 µM. To ensure consistent distribution, 5 µL of RNA will be allotted into each sample's centrifuge tube with stringent mixing protocols. Assuming that all 10,000 unique sequences are amplified at the same rate , the estimated concentration of each individual RNA sequence in the final sample is 225 nM.

At ligand concentrations well exceeding the micromolar range, competition between aptamers has a negligible impact on the effective ligand concentration, making it a suitable threshold for defining saturation conditions. Assuming a hyperbolic relationship between the average ΔSHAPE of an aptamer and the ligand concentration, the fraction of bound aptamers can then be indirectly estimated by normalizing the sequence's average ΔSHAPE reactivity across ligand concentrations, scaling the values from 0 (no ligand) to 1 (saturating ligand conditions). Once the fraction of bound aptamers are determined for each sequence under each condition, I can approximate the effective free ligand concentration:

$$[L]_{\text{free}} = [L]_{\text{total}} - \sum_{i=1}^{10{,}000} \left( f_{\text{bound}}^{(i)} \cdot [A_i] \right)$$

Which can be used to estimate the binding affinity of each aptamer ($i$) :

$$K_d^{(i)} = \frac{[L]_{\text{free}}(1 - f_{\text{bound}}^{(i)})}{f_{\text{bound}}^{(i)}}$$

This ligand-free approximation makes three important assumptions:

1) A hyperbolic relationship between the average ΔSHAPE of an aptamer and the ligand concentration can be used to determine the fraction of aptamers bound at each ligand concentration.
2) Each aptamer binds independently and there is no cooperativity or allosteric effects.
3) The aptamer concentration ([A]) is known and uniform in each sample.

To confirm that condition number 3) is met, I need to ensure that the aptamers themselves do not have subnanomolar binding affinities to each other that would lower the effective aptamer concentration necessary to estimate ligand binding affinity. It would require ~50 million pairwise computations to compute estimated binding affinities of each of the 10,000 RNA sequences to every other sequence. This can be dramatically reduced by first filtering for sequences with more than 15 total complementary nucleotides as well as a scanning for windows of complementary bases that are 7 nucleotides long. If there are sequences that are flagged as potentially being nanomolar binders, I will dilute the initial concentration of RNA in my sample to ensure that the probability of RNAs binding to each other is negligible. Additionally I will use optimized salt concentrations for this protocol to stabilize intramolecular folding over intermolecular interactions.

To validate whether this ligand-free approximation can be used in a competitive aptamer environment to accurately determine binding affinities, I will perform a test experiment by performing this chemical mapping protocol on a library of TPP aptamers and evaluate for correlation with experimentally reported binding affinities.

In summary, this pipeline will produce a total of 2 million reactivity data on unique RNA-small molecule interactions that can be used to estimate 320,000 binding affinities of 10,000 aptamer candidates for 32 different small molecules.

RibonanzaNet-SM: Transfer Learning on RibonanzaNet Foundation Model

Our lab's foundation model RibonanzaNet accurately predicts the chemical probing reactivity profile of a given RNA sequence. This model has been successfully fine tuned to also predict secondary structures and tertiary structure motifs. I plan to finetune the pre-trained RibonanzaNet model to predict changes in RNA chemical reactivity profiles upon incubation with ligands. Instead of taking in one input, the model takes in two inputs, an RNA sequence and a small molecule SMILES string, and produces seven different predicted reactivity vectors. One vector is the absolute reactivity vector of the RNA sequence without small molecule incubation, and the following six vectors are the predicted reactivity differences after incubation with the small molecule under six different concentrations.

The model will run the MolE inference pipeline to convert the small molecule SMILES string into a pretrained, fixed-dimensional ligand embedding. A Transformer Block processes pairwise RNA embeddings by averaging them over nucleotides, generating individual queries for each nucleotide using the ligand embedding as a key. This updates the RNA pairwise embeddings with ligand-dependent information before they pass through the decoder to predict the seven reactivity vectors. I will use a Mean Column Root Square Error (MCRSE) loss function to enforce per-nucleotide consistency of predictions across the seven small molecule concentration conditions. For each sequence, there are 32 training samples, where each sample consists of a tensor containing seven 1-D reactivity vectors under different conditions. With a total of 320,000 training samples, the training process should not exceed 48 hours when using 10x NVIDIA L40s GPUs.

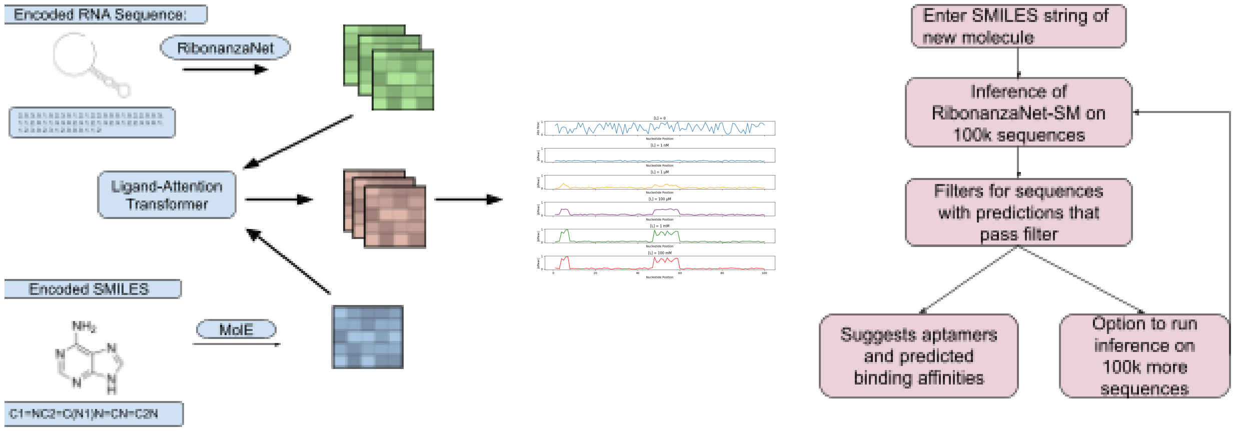
Encoded RNA Sequence:
RibonanzaNet
Ligand-Attention Transformer
Encoded SMILES
MolE
NH2
C1=NC2=C(N1)N=CN=C2N
Enter SMILES string of new molecule
Inference of RibonanzaNet-SM on 100k sequences
Filters for sequences with predictions that pass filter
Suggests aptamers and predicted binding affinities
Option to run inference on 100k more sequences

**Figure 4:** A broad outline of the fine tuned RibonanzaNet-SM architecture that incorporates features from a pre-trained MolE foundation model (left). An example of how the model will be used to practically predict aptamers.

Once the model is trained, I will use it to run inference with a new small molecule. The model will then generate predictions for a user-specified number of RNA transcripts filtered from a database of libraries that include natural and Eterna designed sequences. By default, the pipeline processes 100,000 aptamer candidates, filters for sequences with predicted statistically significant reactivity profiles at the lowest ligand concentration, and identifies those as potential aptamers. I will use a similar measure used by Weeks et al. for RNA targeting drug discovery to filter for statistically significant binding events [6]. A binding event is categorized for a sequence if it has at least three nucleotides with a 20% or greater difference in reactivity from the no-molecule control sample. I will then perform a validation on the predicted aptamers by measuring binding affinities through electrophoretic mobility shift assays (EMSA).

## III. Preliminary Checkpoints

I first wanted to confirm that changes in an RNA's chemical reactivity profile corresponded with small molecule binding using the specific high-throughput protocol developed by my lab to simultaneously probe millions of RNA. A pilot experiment was performed by chemically probing a structurally complex RNA library designed by the Eterna community with seven molecules known to bind highly structured RNA motifs. These modified libraries were simultaneously probed with a standard control sample without small molecule incubation. Resulting analysis of the ΔSHAPE reactivity profiles confirmed that sequences exhibited statistically significant nucleotide reactivity differences that were specific to the ligand identity. This pilot experiment was reproduced by two different members of my lab (including myself) with mitoxantrone, the most nonspecific, promiscuous binding compound from the original pilot experiment. The results were highly reproducible, with a nucleotide resolution level of precision in the corresponding ΔSHAPE reactivity profiles for individual sequences.

Secondly, I wanted to confirm that RibonanzaNet could be fine-tuned to learn chemical reactivity differences. By just modifying RibonanzaNet's decoder to output eight reactivity vectors corresponding to ligand condition for each sequence, the model is able to predict nucleotides with the greatest reactivity differences after just 20 epochs of training. Learning is confirmed by a steadily declining validation loss during training, and performance is evaluated by predictions on a held out test set.

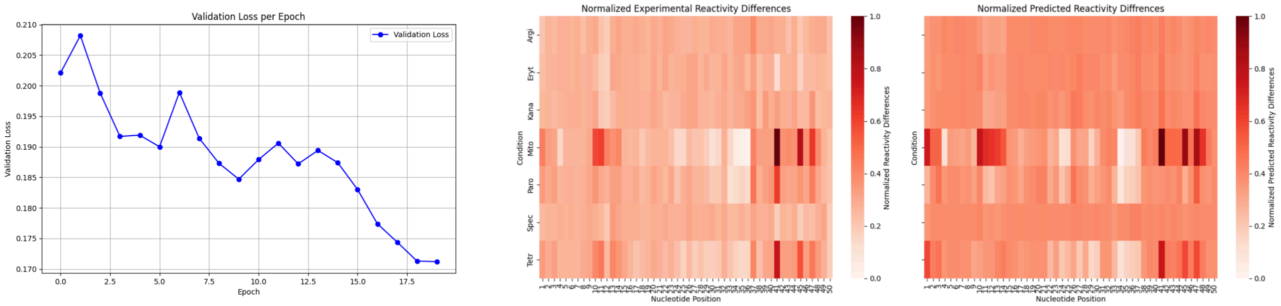

**Figure 5:** Success of fine tuning RibonanzaNet to predict absolute reactivity vectors under eight different conditions can be evaluated based on steadily declining validation loss (left). RibonanzaNet fine tuned on reactivity difference vectors and the absolute reactivity vector of the control assessed by qualitative analysis of the test set (right).

## IV. Future Results/Broader Impacts

Beyond aptamer discovery, developing new experimental protocols to investigate RNA structure and function through its biophysical interactions with small-molecule metabolites is essential for addressing some of biology's most fundamental questions. How did life emerge from an RNA world? Why do humans possess such an abundance of long noncoding RNA? And to what extent do small molecules contribute to stabilizing active RNA structures? These questions drive my curiosity, and I hope this work reveals insight toward answering them as I work toward receiving a doctoral degree.

## References

1) *Wearable healthcare devices market*. MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/wearable-medical-device-market-81753973.html (accessed 2025-02-02).

2) Song, S.; Wang, L.; Li, J.; Fan, C.; Zhao, J. Aptamer-Based Biosensors. *Trends Analyt. Chem.* 2008, *27* (2), 108–117.

3) Yoo, H.; Jo, H.; Oh, S. S. Detection and beyond: Challenges and Advances in Aptamer-Based Biosensors. *Mater. Adv.* 2020, *1* (8), 2663–2687.

4) Tuerk, C.; Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* 1990, *249* (4968), 505–510.

5) Pfeiffer, F.; Mayer, G. Selection and Biosensor Application of Aptamers for Small Molecules. *Front. Chem.* 2016, *4*, 25.

6) Zeller, M. J.; Favorov, O.; Li, K.; Nuthanakanti, A.; Hussein, D.; Michaud, A.; Lafontaine, D. A.; Busan, S.; Serganov, A.; Aubé, J.; Weeks, K. M. SHAPE-Enabled Fragment-Based Ligand Discovery for RNA. *Proc. Natl. Acad. Sci. U. S. A.* 2022, *119* (20), e2122660119.

7) He, S.; Huang, R.; Townley, J.; Kretsch, R. C.; Karagianes, T. G.; Cox, D. B. T.; Blair, H.; Penzar, D.; Vyaltsev, V.; Aristova, E.; Zinkevich, A.; Bakulin, A.; Sohn, H.; Krstevski, D.; Fukui, T.; Tatematsu, F.; Uchida, Y.; Jang, D.; Lee, J. S.; Shieh, R.; Ma, T.; Martynov, E.; Shugaev, M. V.; Bukhari, H. S. T.; Fujikawa, K.; Onodera, K.; Henkel, C.; Ron, S.; Romano, J.; Nicol, J. J.; Nye, G. P.; Wu, Y.; Choe, C.; Reade, W.; Das, R.; Eterna participants. Ribonanza: Deep Learning of RNA Structure through Dual Crowdsourcing. *bioRxiv*, 2024.

8) Brown, A.; Brill, J.; Amini, R.; Nurmi, C.; Li, Y. Development of Better Aptamers: Structured Library Approaches, Selection Methods, and Chemical Modifications. *Angew. Chem. Int. Ed Engl.* 2024, *63* (16), e202318665.

9) Deigan, K. E.; Li, T. W.; Mathews, D. H.; Weeks, K. M. Accurate SHAPE-Directed RNA Structure Determination. *Proc. Natl. Acad. Sci. U. S. A.* 2009, *106* (1), 97–102.