RNA SHAPE PREDICTION

James Swomley, Leon Kloker
Stanford University, Institute for Computational and Mathematical Engineering

ABSTRACT

The ability to easily compute the *selective 2'-hydroxyl acylation analyzed by primer extension* (SHAPE) value associated with each nucleotide in an RNA sequence is an important intermediate step that will allow researchers to then determine the secondary and tertiary structure of a given RNA molecule. This tertiary structure is important as it informs the molecule's function, particularly regarding its transcription of proteins. SHAPE data provides insight into the local flexibility of an RNA molecule at a single-nucleotide resolution which, in turn, informs the folding and spatial arrangement of the RNA molecule. The SHAPE profile of an RNA sequence therefore is expected to contain all of the information necessary to reconstruct the tertiary structure of an RNA molecule. This makes SHAPE a desirable intermediate step between RNA sequence and tertiary structure, as directly computing tertiary structure remains difficult and expensive. A model that can accurately compute SHAPE data could rapidly create a dataset enabling the development of a new model that predicts tertiary structure from SHAPE profile.

This paper evaluates various deep learning models for their efficacy in predicting SHAPE data from RNA sequence alone. Our core dataset consists of 128,400 distinct RNA sequences, curated to ensure a high signal-to-noise ratio. Each sequence in the dataset is associated with two sets of SHAPE data derived from different chemical mapping experiments, providing comprehensive information about nucleotide reactivity. We explore four sequence models: LSTMs, CNNs, GRUs, and transformers, using a multi-head architecture to predict SHAPE values from the different experiments simultaneously. We also introduce a foundation model, RNA-FM, to enhance prediction accuracy by providing information-rich embeddings. The RNA-FM model, inspired by BERT-style architectures, encodes evolutionary and sequential information about RNA and has been pre-trained on a vast corpus of unannotated RNA sequences. We additionally attempt to fine-tune select layers of the RNA-FM model in order to improve prediction accuracy. A multi-task architecture with an additional prediction head to output secondary structure of the given RNA sequence is also considered in order to introduce some inductive bias to force the model to focus on structural awareness.

The GRU-based models outperform all other trialed sequence models. The success of the relatively simple GRU suggests the importance of short-range sequence dependencies in determining the SHAPE value of a nucleotide. The use of RNA-FM embeddings proves substantially helpful in improving model accuracy over simple sequence model architectures, with all RNA-FM supported models outperforming all basic sequence models. The multi-head architecture incorporating both types of SHAPE values also improves model functionality without sacrificing accuracy. Fine-tuning was not successful in our trials. Ultimately, a GRU trained on RNA-FM embeddings with two prediction heads for the two SHAPE value types performs the best, with a 0.751 median correlation coefficient between experimental and predicted SHAPE profiles.

In conclusion, our study demonstrates the feasibility of using deep learning models, especially those incorporating RNA-FM embeddings, to predict SHAPE data from RNA sequences. This approach not only provides a cost-effective alternative to experimental methods but also opens new avenues for computational RNA tertiary structure prediction. By leveraging machine learning, we move closer to bridging the gap between RNA sequencing and structure determination, enhancing our understanding of RNA functionality and its role in biological processes.

1 Introduction

The secondary and tertiary structure of an RNA molecule, as shown in Figure 1, are crucial to its transcription of proteins, since they determine the reactivity of its individual nucleotides depending on whether they are paired or unpaired. However, directly measuring the structure of individual RNA molecules remains difficult, even with modern methods such as cryo-electron microscopy (cryo-EM), which is in direct contrast to the relative triviality of simply sequencing RNA molecules. The ability to infer such structures from only the sequence of the RNA molecule would help bridge this gap.

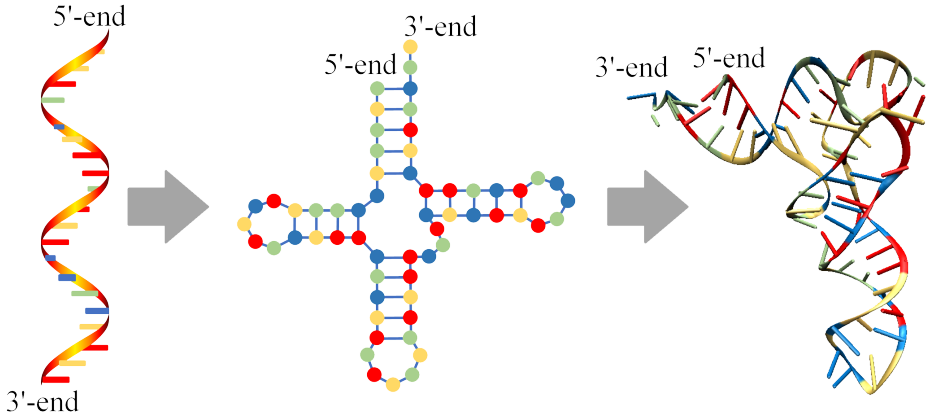

Figure 1: An RNA sequence, its secondary structure, and its tertiary structure.

An intermediate step in this process is given by the *selective 2'-hydroxyl acylation analyzed by primer extension* (SHAPE) data associated with an RNA molecule. This measures the local reactivity of each nucleotide in the molecule with a certain reagent, such as 1M6, 1M7, or NMIA [1], and in turn informs the local flexibility of the RNA at a single-nucleotide resolution, with paired nucleotides exhibiting a lower flexibility in most cases than their unpaired counterparts.

This means that the SHAPE data is of use in determining the secondary structure of an RNA molecule, which is done in various papers such as Spasic et. al. [2]. It is also expected to completely encode the tertiary structure, in that a perfect reconstruction of the SHAPE data of a molecule will allow for the computation of its tertiary structure. Given that the experimental determination of the tertiary structure through cryo-EM comes in at around $1,000 per molecule, whereas extraction of SHAPE data costs only $1 per molecule, it is much more amenable to machine learning methods at the moment, and indeed Stanford's Das Lab has recently released an appreciable amount of SHAPE-labeled RNA sequences. However, the task of accurately calculating the SHAPE values for a given sequence of nucleotides remains largely open and is what our research addresses.

In this paper, we evaluate a number of deep learning models trained to predict SHAPE data from RNA sequences, finding a model centered around a foundation model and gated recurrent unit (GRU) to perform best.

2 Data

The data used to train the models consists of 115,200 distinct RNA sequences ranging from length 115 to 206, specifically selected from a larger set of sequences to ensure high signal-to-noise ratio and read coverage. In addition to the training data, a validation and test dataset both containing 6,060 RNA sequences, which amounts to 5% of the overall data, were used. The four different nucleotides of the RNA sequences are represented using simple integer numbering.

Each sequence comes with two sets of SHAPE data for the two different types of chemical mapping experiments used to generate the profiles: one using dimethyl sulfate (DMS) [3] and one using 2-aminopyridine-3-carboxylic acid imidazolide (2A3) [4]. The two sets of SHAPE profiles show significant, but not complete, agreement. The SHAPE profiles, consisting of a single scalar value for each nucleotide, are clipped to fit between zero and one as larger outliers arise from measurement artifacts and are not representative of actual reactivity [5].

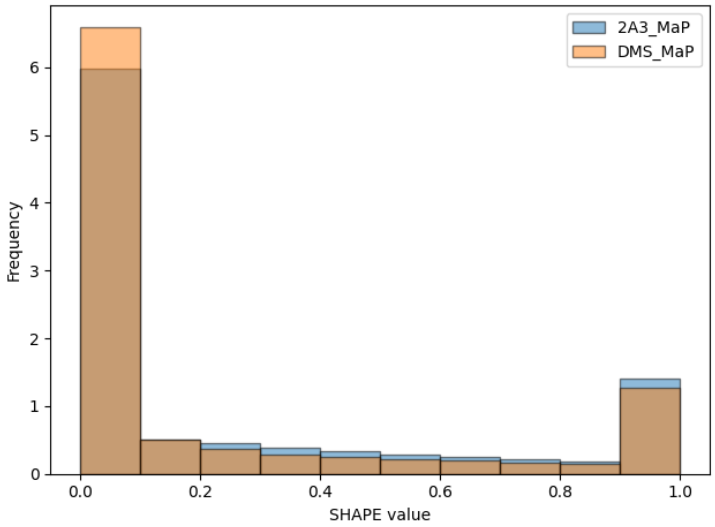

Figure 2: Density histogram of the SHAPE values in the training dataset.

The dataset also contains the secondary structure for each sequence in dot-bracket format, showing which nucleotides are bound to which other nucleotides. For our use case, the dot-bracket notation was transformed to several different forms of data as later explained in section 3.4.

3 Modeling Framework

A general outline of the various modeling frameworks is provided below in Figure 3.

3.1 Baseline

Our baseline model follows the red path in Figure 3. Each RNA sequence is embedded in a 256-dimensional space using PyTorch's `nn.Embedding` function, which does so by combining a lookup table with a learned set of weights. The embedding is then passed to a sequential model, of which we tested four: an LSTM, a GRU, a CNN, and a transformer encoder. Throughout the layers of each of the sequence models, the dimension of the nucleotide embeddings are maintained. Finally, a 3-layer fully-connected prediction head was used to transform the 256-dimensional encoding of each nucleotide in the sequence into one scalar value. Separate prediction heads for 2A3 and DMS SHAPE values are used that both operate on the output of the preceding sequence model. As both SHAPE profiles are similar, using two prediction heads with the same core model should provide more information during training, therefore creating a more robust model. The loss function weights both predictions equally.

3.2 Foundation Model

To improve and augment the information our model has access to, we introduce a foundation model, RNA-FM [6]. RNA-FM is a BERT-style self-supervised foundation model consisting of twelve transformer encoder layers trained on a massive corpus of unannotated RNA sequences by masking 15% of the nucleotides and then predicting the masked tokens based on the 640-dimensional nucleotide encoding. The model is designed to learn and encode evolutionary and sequential information about RNA, similar to what ESM-2 [7] does for proteins.

To incorporate RNA-FM, instead of using a general purpose embedding and passing the RNA sequence directly into the prediction head, we follow the orange path in Figure 3 and pass our RNA sequence first through RNA-FM. RNA-FM then outputs information-rich embeddings for the RNA sequences, upon which we train our sequence model. Here, the sequence models are fed the 640-dimensional RNA-FM encoding but still operate on a 256-dimensional space throughout their layers. Prepending the foundation model should improve the accuracy of our SHAPE predictions, as the model has access to significantly more information than is present in our SHAPE-labeled dataset and the RNA-FM encoding has been proven to improve performance on downstream tasks [6].

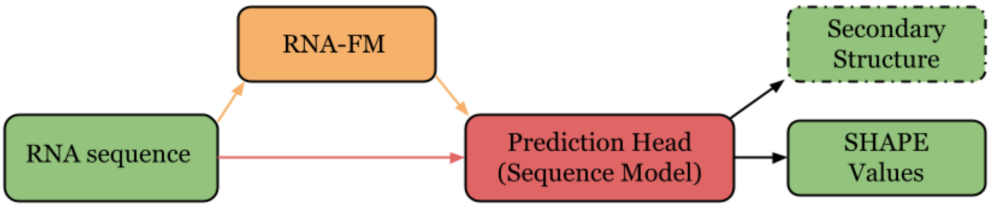

Figure 3: Flowchart outline of the modeling framework.

3.3 Fine-Tuning

A more expressive way to employ RNA-FM involves fine-tuning. This is accomplished by initializing the RNA-FM model with its pre-trained parameters and then retraining it for our specific task. This allows the model to remain similar to its initial state, accomplishing the same task it was originally trained to do, but slightly optimized and adapted to fit our current task.

In our case, we prepend the RNA-FM model to our sequence model and then train them together, freezing some layers of RNA-FM to prevent them from being updated. As the memory of our GPU on Azure is limited to 8GB, we apply fine-tuning only to the last 3 transformer encoder blocks as unfreezing more layers while maintaining a reasonable batch size leads to memory issues.

3.4 Multi-Task Learning

An additional way to incorporate more information into our model involves the introduction of a third prediction head. As mentioned in section 2, in addition to the two types of SHAPE values, we also have access to secondary structure in the form of nucleotide binding information. Given the high correlation between the SHAPE values and the structural information of each sequence as mentioned in section 1, the structural awareness of each model is increased by training on this additional output and we hope to introduce some inductive bias that helps the model to generalize better [8]. The prediction of various nucleotide-level structural metrics were trialed, including a simple binary hydrogen-binding indicator, and a variety of graph metrics including load and closeness centrality of each nucleotide as a node in the RNA connectivity graph.

3.5 Hyperparameters

The models were trained using AdamW with a one-cycle learning rate scheduler that anneals the learning rate until 10% of the training is finished. The start, max and min learning rates are 1e-3, 5e-3 and 5e-5, respectively. The learning rates for fine-tuning are chosen as 1e-5, 5e-5 and 1e-6 in similar order. The loss function is given by the mean absolute error (MAE) of both the 2A3 and DMS SHAPE prediction, which are weighted equally. For predicting secondary structure in the form of a binary indicator predicting nucleotide binding, we utilize binary cross entropy as loss which is weighted by a factor of 0.25. Moreover, training is done for 100 epochs with a batch size of 256 in all cases except for fine-tuning. Here, due to GPU memory constraints, we utilize a batch size of 32.

As mentioned previously, the sequence models use a hidden dimension of 256. For the GRU and LSTM, this means that the hidden dimension as defined in PyTorch is 128 since they operate bidirectional which produces a double-sized hidden state. The transformer encoder is also used bidirectionally. Furthermore, we found 3 layers for all models to be a good trade-off between model capacity and ease of the learning procedure. In this configuration, all models including their fully-connected prediction heads have a size of around 1.5 to 2 million parameters.

4 Results

In addition to the loss (MAE), the performance of the model is also evaluated by the median of the Pearson correlation coefficients, defined sequence-wise as

$$r_{seq} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i^* - \bar{y}^*)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2}}, \tag{1}$$

where y_i and y_i^* are the experimental and predicted SHAPE values at nucleotide i of the sampled sequence, respectively, and $\bar{y}$ and $\bar{y}^*$ denote their mean values.

Our primary set of trials involve comparing the baseline sequence model with the sequence model trained on RNA-FM embeddings, for a variety of sequence model types. These models all use two prediction heads, predicting the SHAPE values corresponding to DMS and 2A3. All performance metrics shown are calculated considering all results across both heads.

method	metric	CNN	Transformer	LSTM	GRU
baseline	MAE	.204	.391	.295	**.190**
	correlation	.594	.165	.278	**.637**
w/ RNA-FM embeddings	MAE	.177	.195	.169	**.162**
	correlation	.707	.623	.735	**.751**

Table 1: Model Performance Metrics on Test Data

The GRU proves to be the best performing sequence model in all cases, considering both Pearson correlation coefficient and mean absolute error. The CNN and LSTM both perform somewhat worse than the GRU in all cases, while the transformer underperforms all of them. In the baseline model, the transformer performs very poorly despite being the most state-of-the-art and expressive sequence model out of the four tested models.

The use of RNA-FM embeddings improves model performance across the board. The most drastic improvements are exhibited when the embeddings are prepended to initially poor performing baseline models, like the transformer and LSTM. Using RNA-FM embeddings levels the discrepancy between each sequence model's performance, shrinking the difference in correlation coefficient between best and worst from 0.472 to 0.128.

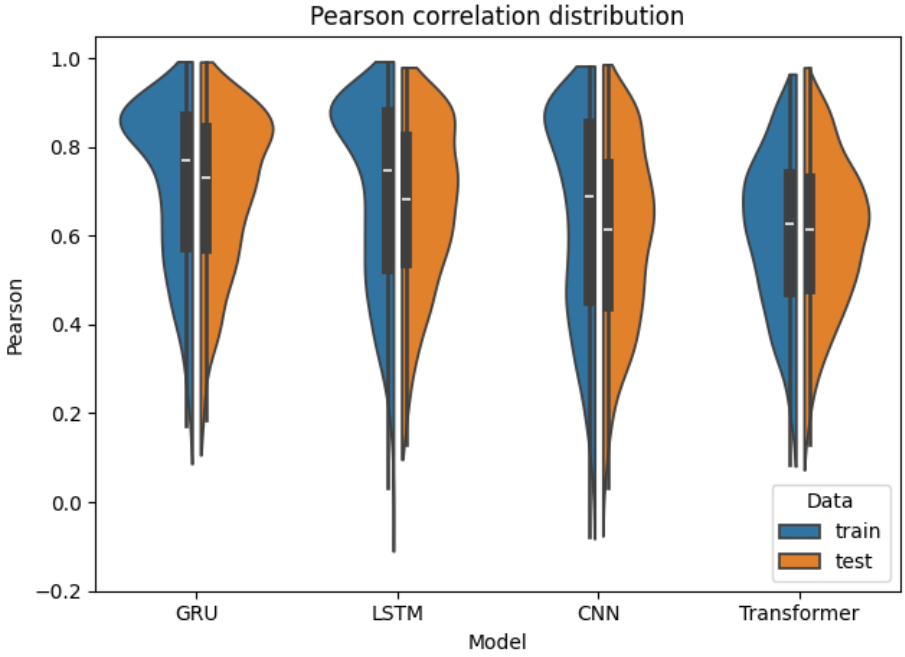

Figure 4: Distribution of correlations across supervised models.

4.1 Multi-Task Learning

In terms of multi-task learning, the use of two prediction heads for DMS and 2A3 SHAPE prediction was immediately adopted to improve the model's functionality. The aggregate performance across both heads is consistently equal to or slightly better than the performance of a single-headed model predicting only one type of SHAPE value, so there is no sacrifice in performance associated with the improved functionality of the multi-task structure, and therefore no reason for it not to be utilized.

Adopting a third prediction head to incorporate structural information was not as beneficial. The most promising implementation was predicting a binary value for each nucleotide corresponding to whether it was bound or unbound.

While this did slightly improve the correlation coefficient of some of the worst models, like the baseline transformer and LSTM, its effect was unnoticeable for any of the models trained on RNA-FM embeddings. Additionally, more informative types of structural information were trialed without noticeable improvement. By constructing RNA connectivity graphs based on the molecule's secondary structure, node/nucleotide-level graph measures like load and closeness centrality can be extracted. The prediction of these graph measures yielded worse results than the predicting the binary values, hurting accuracy in most cases.

4.2 Fine-Tuning

Fine-tuning the last three transformer-encoder layers of RNA-FM with a GRU appended yielded a Pearson correlation coefficient of 0.724 and a mean absolute error of 0.176. This is a decrease in correlation of 0.027 and an increase in MAE of 0.014 from the same model without fine-tuning.

5 Discussion

Historically, not much work has been completed in the way of RNA SHAPE prediction, due to a lack of data. Previous attempts to tackle the problem, like Bliss et. al.[9], had to operate with far less data. In Bliss et. al. they used a CNN with attention far larger than the models presented in this project, achieving a median correlation of about 0.5 while training and testing on only 194 and 32 RNA sequences, respectively. While technically completing the same task, the difference in data availability transforms the project, making any model comparison essentially irrelevant. Due to the recent release of a high volume of SHAPE annotated RNA sequences by Stanford's Das Lab, many more attempts at this task are being completed. For the time being, however, the entirety of this project is novel and original.

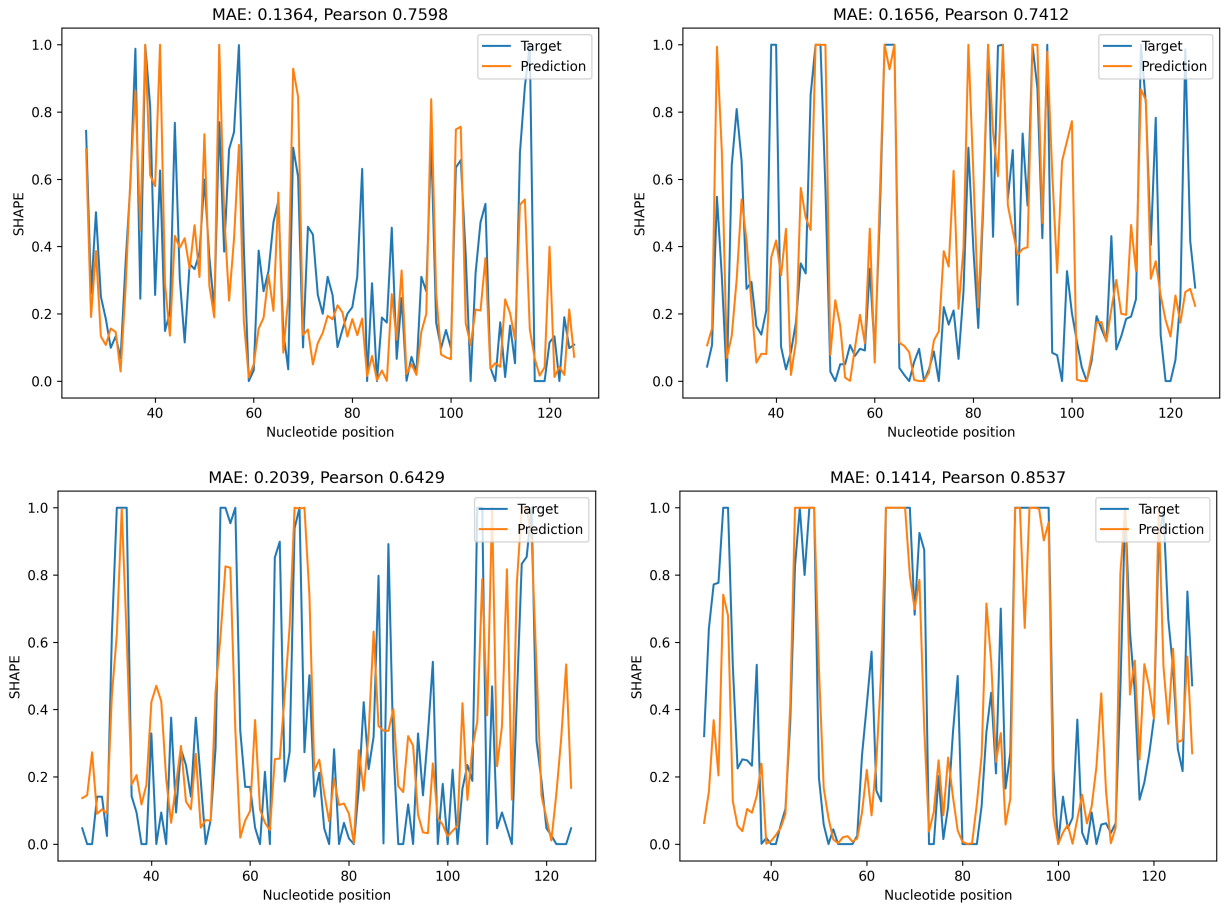

Figure 5: Four example SHAPE profile predictions from the GRU with RNA-FM embeddings model on the test dataset.

The GRU models outperforming all other tested sequence models is a surprising result, as the GRU is the most simple of the tested models except for the CNN, especially in comparison to the transformer. Since the GRU and LSTM are similar, both being types of recurrent neural networks, the fact that the GRU outperforms the LSTM can tell us about the nature of SHAPE data. Since an LSTM is better-suited for capturing long-term dependencies, it is likely that much of the information that informs nucleotide reactivity and therefore SHAPE is found close to that nucleotide. Focusing on shorter-term dependencies may well be an advantage for the GRU. The transformer, which is designed to process a wider range of data before finding dependencies, may perform relatively poorly due to an inability to immediately hone in on the relevant short-term dependencies the way the GRU inherently does. Since transformers are more powerful overall, they have the potential to outperform the other options if trained or appropriately fine-tuned better. However, in our conditions, we were not able to train the best transformer model.

Our best performing model, the GRU trained on RNA-FM embeddings, exhibits interesting behavior. As shown in Figure 5 below, the model has little issue predicting the location of peaks in the data. These peaks correspond to nucleotides that are prone to binding, which is the information that is central to eventual tertiary structure prediction. The prediction error is largely due to an inability to predict the exact magnitude of these peaks and troughs, not an inability to predict their location. This is promising, as it indicates that even a fairly imperfect model can still generate very useful SHAPE data. Additionally, we must consider the optimal Bayes error of the SHAPE dataset, much of which likely comes from unpredictable variations in SHAPE magnitude at the peaks and troughs, especially in SHAPE profiles with high frequency fluctuations as shown in the bottom left graph of Figure 5.

Fine-tuning RNA-FM unfortunately did not improve the model. Fine-tuning a very large foundation model is not only a computationally demanding task but a suboptimal choice of hyperparameters such as learning rate or schedule might also heavily compromise the results. Similarly, choosing which layers of the model to fine-tune can influence the performance significantly. With our choices of hyperparameters, we tried to orient ourselves along the lines of some popular fine-tuning frameworks, especially for BERT since our model is basically similar in architecture. Due to time and resource constraints, however, we were unable to optimize by iterating on the large set of possible hyperparameter combinations.

We expect that for an optimal choice of fine-tuning hyperparameters, the performance of the overall model should increase as the task it was trained to solve is highly related and we already realize a large performance gain just by using the foundation model's embeddings. It might be the case that fine-tuning the first or middle layers of the transformer encoder leads to improvements. Moreover, a more surgical approach to fine-tuning through low-rank adaptation [10] could also be beneficial as this is commonly used for large language model, consisting also of a transformer en- and decoder.

The addition of a structural output prediction head to the model did not affect SHAPE prediction accuracy when using a binary bound/unbound variable, and hurt accuracy when using a more informative graph metric. This is most likely because these two metrics are not that well-correlated with the SHAPE values themselves, even if they do provide some extra structural awareness to the model. Nevertheless, including secondary structure as an additional prediction task can improve the generalizability of a model if used on out-of-distribution sequences that potentially vary significantly in length as the inductive bias that is introduced works as a regularization mechanism.

6 Conclusion

The performance of the baseline models presented in this report, especially the GRU, demonstrate the capability of deep learning models to correctly predict SHAPE values from nothing but the RNA sequence itself. Moreover, making use of the sheer amount of available unannotated RNA sequences through a foundation model that is pre-trained in a self-supervised fashion proves to have an edge over immediate modeling approaches. Directly fine-tuning parts of the foundation model turns out to be a fairly challenging optimization problem and did not lead to improved prediction in our experiments. However, we are still convinced that more surgical fine-tuning approaches such as LoRA could potentially lead to some additional performance gains.

Furthermore, forcing models to incorporate structural features by hard parameter sharing with a secondary structure prediction head does neither significantly improve nor deteriorate performance but could lead to an improved generalization capability when considering sequences of widely varying lengths. Trying to incorporate more sophisticated forms of secondary structure information such as the base-pair binding probability matrix into the modeling pipeline will most likely boost performance in future works. However, in its current state, the inability of any model to accurately predict the height of all SHAPE peaks, as well as the large variance in performance exhibited by the wide spread of Pearson coefficients observed across the test dataset indicates that there remains some room for improvement.

7 Appendix

James Swomley: lead in multi-task model development, data cleaning, model training

Leon Kloker: lead in sequence model implementation, RNA-FM inclusion, fine-tuning

Github: https://github.com/leonkloker/RNA

References

[1] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols*, 1(3):1610–1616, August 2006.

[2] Aleksandar Spasic, Sarah M Assmann, Philip C Bevilacqua, and David H Mathews. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research*, 46(1):314–323, November 2017.

[3] III Mitchell, David, Jennifer Cotter, Irfana Saleem, and Anthony M Mustoe. Mutation signature filtering enables high-fidelity RNA structure probing at all four nucleobases with DMS. *Nucleic Acids Research*, 51(16):8744–8757, 06 2023.

[4] Tycho Marinus, Adam B Fessler, Craig A Ogle, and Danny Incarnato. A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Research*, 49(6):e34–e34, 01 2021.

[5] Rui Huang Jill Townley Rachael Kretsch Thomas Karagianes John Nicol Grace Nye Christian Choe Jonathan Romano Maggie Demkin Walter Reade Rhiju Das, Shujun He and Eterna players. Stanford ribonanza rna folding, 2023.

[6] Jiayang Chen and et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pages 2022–08, 2022.

[7] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[8] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[9] Noah Bliss, Eckart Bindewald, and Bruce A. Shapiro. Predicting RNA SHAPE scores with deep learning. *RNA Biology*, 17(9):1324–1330, May 2020.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.