Capstone Final Report

Using transformer to extract structure information for multi-sequences to boost function learning

Aihan Liu, Hsueh-Yi Lu

Master of Data Science, The George Washington University

Capstone_DATS 6501_81

Dr. Edwin Lo

May 8, 2023

Table of Contents

Abstract:

This project focuses on using transformers to extract structural information for multi-sequences, which enhances function learning in RNA. RNA utilizes four bases, and nucleotide-nucleotide interactions are crucial in determining RNA structure and function. Transfer learning from protein-based models can improve RNA contact prediction and deep learning algorithms can be used to identify and prioritize important positions in genotype-phenotype associations. The study also explores graph convolutional networks to learn complex structure-function relationships. The results indicate that machine learning-based structure and function prediction can provide explanations for RNA structure and open up avenues for new research in the field of molecular biology.

1. Introduction:

Machine learning-based biocomputing methods have made breakthroughs in the field of proteins. Proteins are only one kind of biomolecule, and genes (DNA/RNA), as the source of proteins, contain more basic information and have more critical research value than the latter. The main reason is the lack of structural details because structure determines function; knowing the structure is necessary for the function to be inferred.

RNA structure and function are essential for a wide range of biological processes, including gene expression, regulation, and catalysis. Understanding the relationship between RNA sequence, structure, and function is a fundamental problem in molecular biology. However, predicting RNA structure and function remains a challenging task due to the complexity of RNA molecules and the limited availability of experimental data.

RNA structure is determined by the interactions between nucleotides, including base pairing and stacking, and tertiary interactions, such as loop interactions and base triples. These interactions play a critical role in determining the stability and function of RNA molecules. Thus, accurate prediction of RNA structure requires a comprehensive understanding of these interactions.

In recent years, deep learning-based methods have shown great promise in predicting RNA structure and function. For example, the coevolution-Transformer transfer learning model (Jian et al, 2023), a protein-based model, has been used to predict RNA contact, which is critical in determining RNA structure and function. This approach has significantly reduced the data scarcity bottleneck for RNA structural prediction.

In addition, machine learning-based approaches, such as deepBreaks (Rahnavard et al., 2023), have been developed to identify and prioritize important positions in genotype-phenotype associations. By fitting multiple models and selecting the best one based on cross-validation score, these methods can predict the phenotype based on the provided sequence and interpret it to find the most discriminative positions.

Moreover, graph convolutional networks (GCNs) (Kipf et al., 2016), a neural network designed to work with graph-structured data, have been shown to be effective in learning and predicting tasks on graphs. By using a convolutional operation to aggregate information from a node's local neighborhood, GCNs can effectively capture the complex interactions between RNA nucleotides and learn structure-function relationships.

2. Methodology

2.1. Dataset

RNA is composed of four bases: A, U, C, and G, and RNA sequences are combinations of these nucleotides. Tertiary nucleotide-nucleotide interactions are essential for determining RNA structure and function, and physical distances of less than 10 Å are considered "contacts."

This project used two types of dataset:

- **For structure prediction:** RNA data to train(57)/test(23) transformers.
- **For function learning:** RNA data with 6 different diseases. Each disease has a different number of sequences with different functions.

The RNA sequences are saved into different files, which represent different RNA. Inside each file, There are multi-sequences of RNA.

Fig1: RNA data details

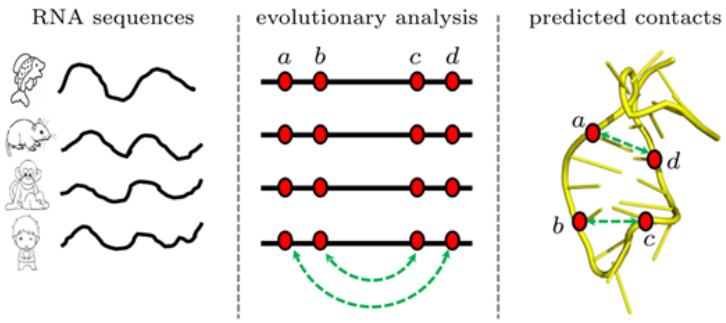

Fig2: RNA data structure based on co-evolution across species

The multi-sequences are the same RNA from different species for RNA of structure prediction. As Fig2 shows, the different sequences have the same structure based on the co-evolution across species.

For RNA of function learning, multi-sequences inside each RNA file vary in their functions. For example, there is an RNA file for the SARS-CoV-2 virus, which could be classified into Alpha and Delta variants.

2.2. Methods

In this study, our group proposes a three-stage method to predict RNA contact using the coevolution-Transformer model and investigate the number and distribution of intersecting connections (cuts) in RNA structure. Our group also uses deepBreaks to predict the function of RNA and explore the feature importance to provide explanations for the results from the coevolution-Transformer model. Finally, our group implements a GCN with two graph convolutional layers to learn complex structure-function relationships. Our approach is promising to advance our understanding of RNA structure and function. It can potentially lead to new discoveries and breakthroughs in molecular biology.

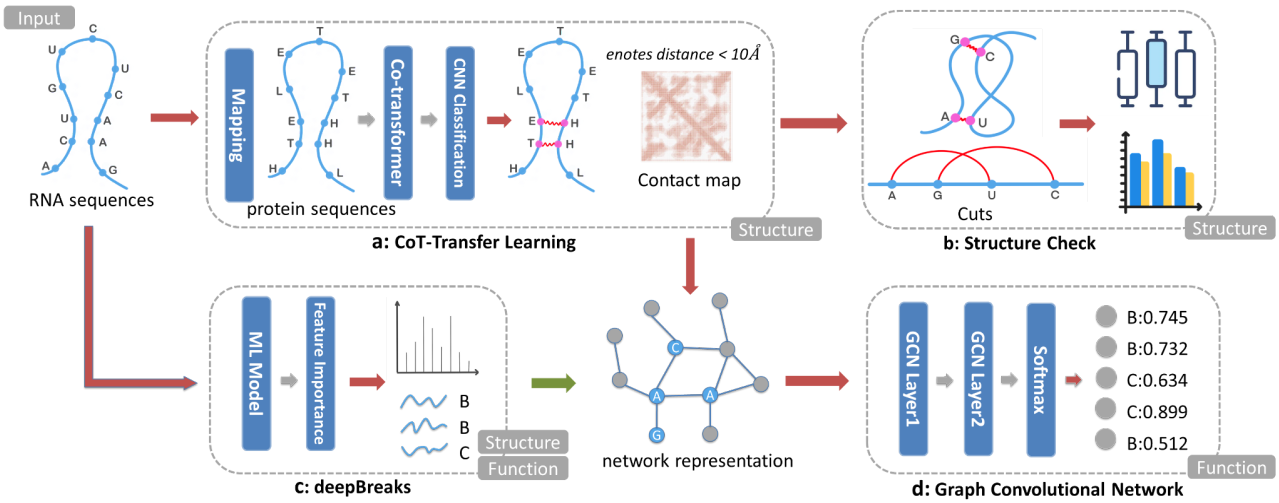

Fig3: Workflow for this project. **a:** A three-stage method to predict the RNA contact based on the protein co-transformer model. **b:** Study of number and distribution for cuts from the prediction based on Co-Transferred result. **c:** Used the deepBreaks to predict the function for RNA. The most important features could overlap with contact maps, which provide an explainable result for RNA structure. **d:** A GCN with two graph convolutional layers for learning complex structure–function relationships.

CoT-Transfer Learning is a method of transferring structural patterns from proteins to RNAs, improving RNA contact prediction and overcoming data scarcity in RNA structural prediction.

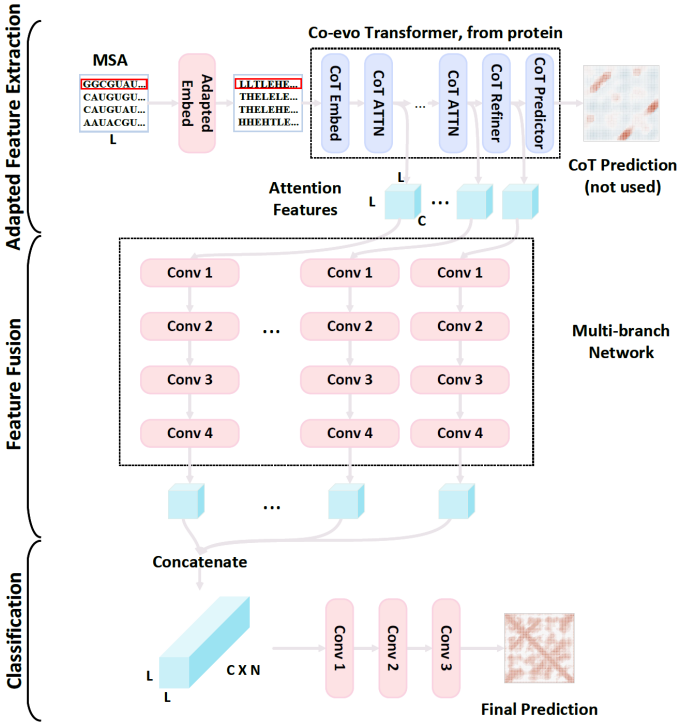

Fig4: CoT-Transfer Learning structure

The three-stage method (from top to bottom) shows in **Fig4**. The Adapted Feature Extraction stage uses a projection layer to translate the RNA MSA sequences into protein language (e.g., from nucleotide “AUCG” to amino acids “HETL”). Then applied these transferred sequences into the co-evolution transformer model (CoT), which uses a pre-trained protein dataset to extract the attention features. The feature Fusion stage uses convolution blocks to concatenate the features from different attention layers. The classification stage aggregates the features into a standard convolutional classifier.

Structure check is applied after the structure prediction. In order to analyze the contact patterns of a protein, a crucial concept to understand is the definition of a "cut." A cut refers to the breaking of contact between two specific points on a protein, and is determined based on whether or not a third point falls within the range of the initial two points. Specifically, a cut is identified when there is contact between point i and j, and there is also contact between point k and w. If either k or w falls within the range of i and j (i<= k <= j) or (i<= w <= j), then a cut is considered to have occurred. This definition allows for a precise understanding of the contact patterns within a protein and can aid in identifying key regions or structures.

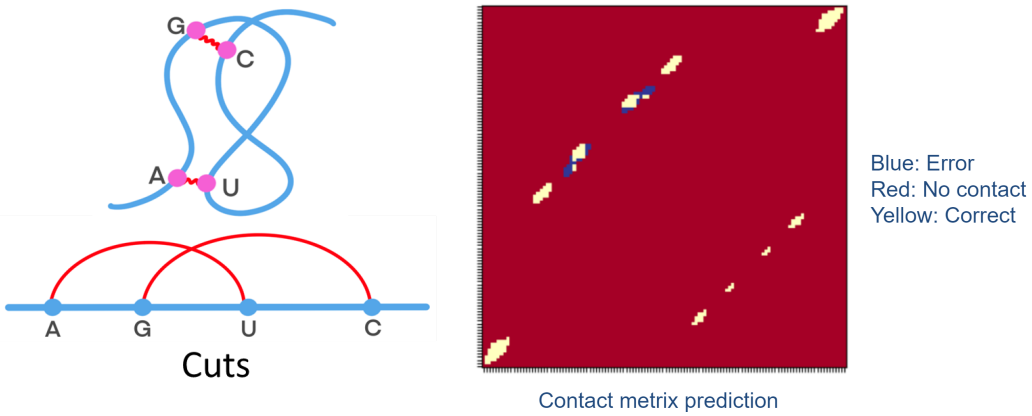

Fig5: Definition of Cuts

DeepBreaks is a computational method that aims to identify significant changes associated with a phenotype of interest using multi-alignment sequencing data from a population. The method involves transferring the data into a position-based matrix and fitting it into different machine learning models. Subsequently, the results are compared to determine the model with the highest accuracy, which is then used to determine the important positions or points of the RNA sequence. This process helps to identify genomic regions and genetic variants significantly associated with phenotypes of interest. The methodology adopted in deepBreaks prioritizes statistically promising candidate mutations, thereby reducing the computational burden associated with

checking all possible mutations. The software is user-friendly, open-source, and comes with high-quality visualization and statistical testing capabilities, making it ideal for researchers working in the biology field. Furthermore, deepBreaks has been benchmarked for computing time to handle high volume sequence data, making it suitable for use in real-world applications.

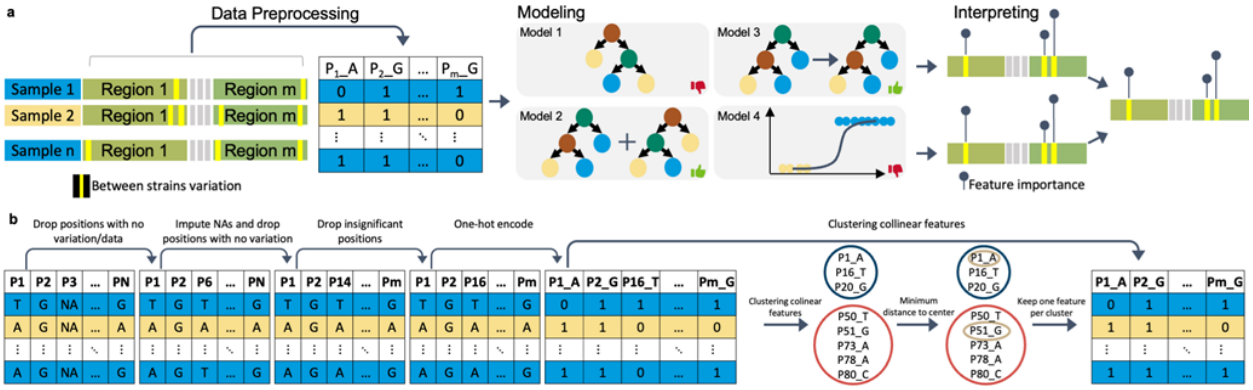

Fig6: deepBreaks

Graph Convolutional Networks (GCNs) is a type of neural network designed to work with graph-structured data, such as social networks or molecular structures. A graph is a collection of nodes connected by link, and GCNs use a specialized convolution operation to learn and propagate information across the graph. This allows GCNs to capture both local and global patterns in the data, and to make predictions about nodes that have not been seen before. GCNs have become popular in recent years due to their ability to learn from complex and especially structured data.

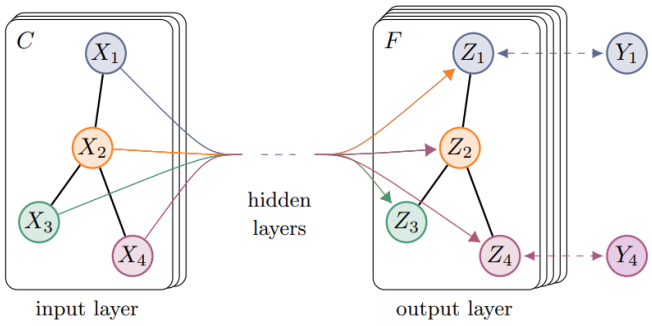

Fig7: Graph Convolutional Network

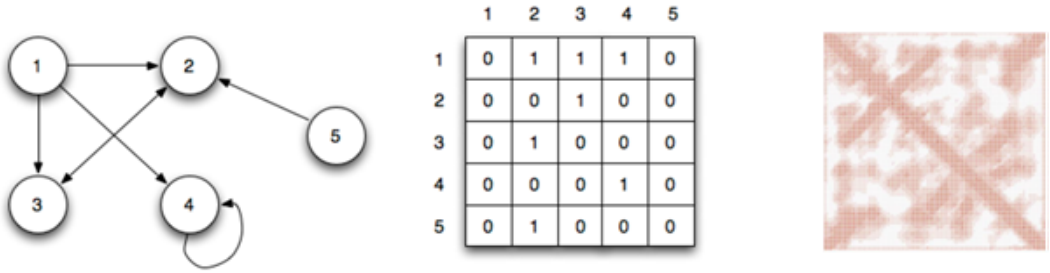

Fig8: Adjacency Matrix and Contact Map

The input of GCN should include the nodes and links. In this case, the input for nodes is different nucleotides, while the links are the adjacency matrix. The adjacency matrix is a matrix with rows and columns labeled by graph vertices, with a 1 or 0 according to whether these two positions are adjacent or not. A contact map is an adjacency matrix for RNA.

3. Results:

3.1. Cut numbers

Based on the analysis of the distribution plots, there is no significant difference between the distribution of the number of error cuts and the correct cuts. The plots show no distinct pattern or trend, indicating a noticeable distinction between the two distributions. Therefore, it can be inferred that the error and correct cuts have a similar distribution, and their frequencies are distributed evenly without any significant bias. These findings suggest that the error cuts occur randomly, without any apparent pattern or correlation, so it is challenging to predict their occurrence accurately.

More results could be found in Appendix A and Appendix B.

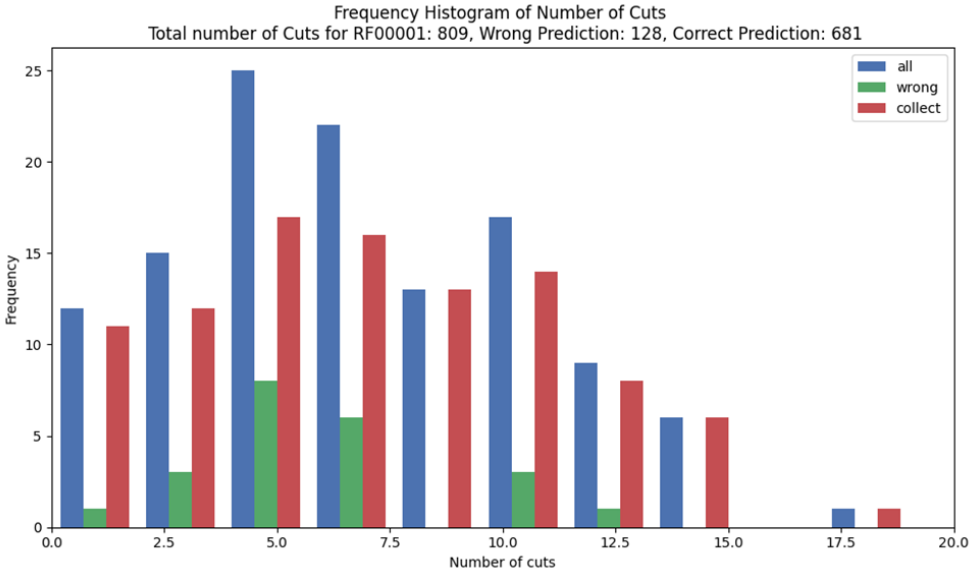

Fig9: Distribution for cuts for RF00001

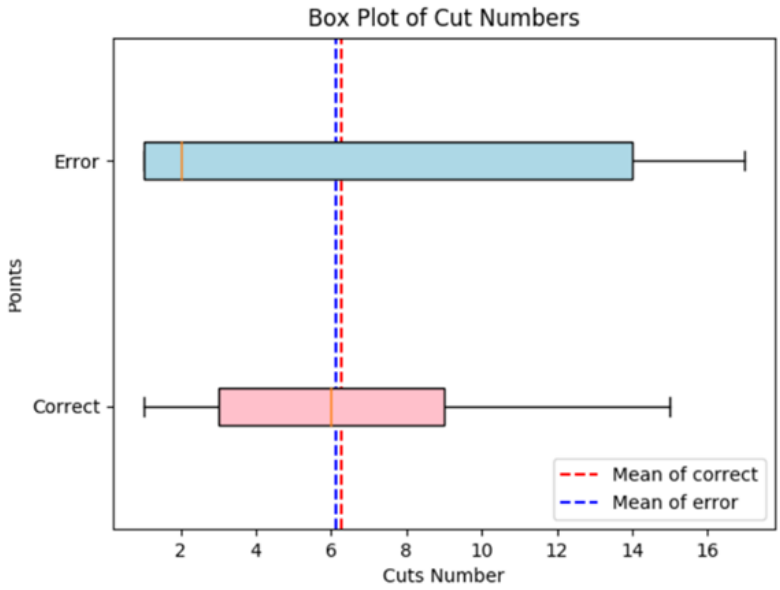

Fig10: Boxplot for cuts for RF00001

3.2. deepBreaks

The below result shows the deepBreaks result using HIV data. The prediction achieved 99.17% accuracy for the XGBoost Classifier (with default parameters).

In analyzing a dataset, the position importance was generated in Fig11, showing that certain positions are significantly more relevant than others.

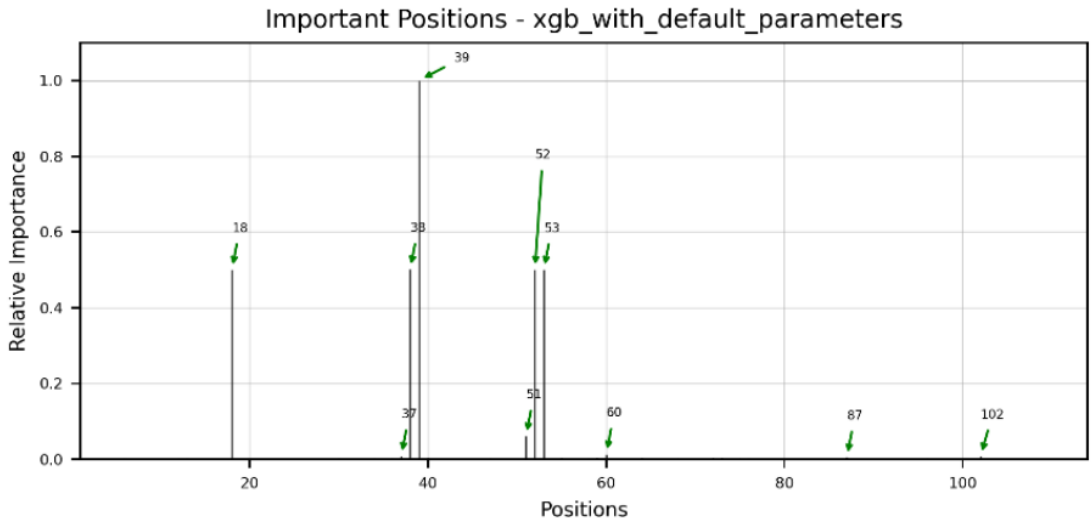

Fig11: Position of Importance for HIV data

If compared these important positions with the contact predictions, it can be observed that some of these points match the contact matrix. In Fig12, the yellow points are the

contact predictions, while the red points in the figure are the points that match the significant locations.

This could validate the prediction of an important position from deepBreaks and also confirm the important position and the closer points should have some internal similarity.

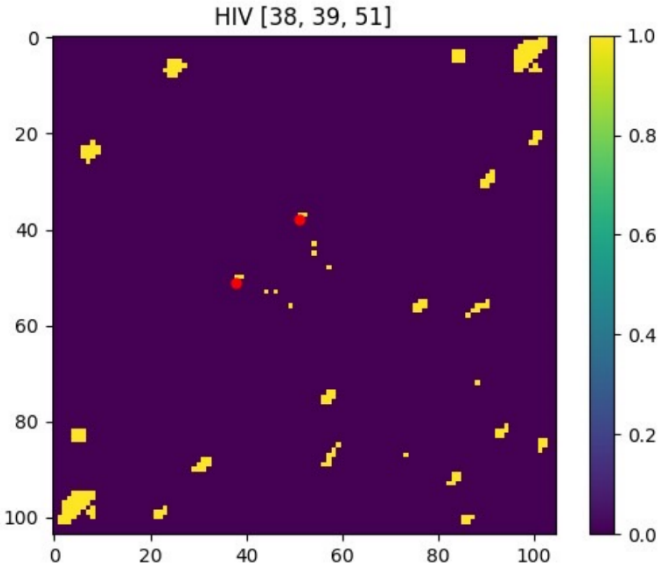

Fig12: Overlap Result of the important position and contact predictions

3.3. GCN

To confirm whether adding contact information will lead to better predictions, multiple experiments were applied for GCN. The HIV data has 35,000+ sequences with 105 nucleotides (imagine it as 35k sentences with 105 words on each), which means the original data set has the shape of (35,000, 105). And the SARS-CoV-2 data has 900 sequences with 3822 nucleotides. Since the sequence is too long to leverage into the CoT-Transfer Learning model, the first 500 nucleotides are taken for the contact prediction.

- Assigned a random contact matrix to the GCN model with the original dataset. The prediction results are similar to each other (shows green in Fig13).
- Reduced the training size to 100, with the random contact matrix, the accuracy is 80%, while if used the real contact matrix, the accuracy is 100%.
- Applied the GCN model to SARS data. The result with real contact has 95% accuracy , and the accuracy for random contact is slightly lower (93%).
- Reduced the training size to 300. The random contact had an accuracy of 0.34 , while the real contact had an accuracy of 0.91 .

Data Description	Model Name	Training Size	Accuracy
HIV-1 based on V3: 35424 sequences 105 nucleotides 2 classes	*Logistic regression	35,424, 105	0.9925
	GCN, real contact	35,424, 105	0.9924
	GCN, random contact	35,424, 105	0.9883
	GCN, real contact	100, 105	1.0000
	GCN, random contact	100, 105	0.8000
SARS-CoV-2: 900 sequences 3822 nucleotides 2 classes	*Extra Trees Classifier	900, 3822	0.9745
	GCN, real contact	900, 500	0.9495
	GCN, random contact	900, 500	0.9376
	GCN, real contact	300, 500	0.9184
	GCN, random contact	300, 500	0.3469

* Result from the best ML method provided by deepBreaks

Fig13: Results for GCN with /without structure information

Even though there are only two RNA examples to compare, it still could be concluded that with large samples for training, the contact doesn't help increase prediction accuracy and the type of nucleotides is more critical in this case. While if the number of samples is smaller, the contact significantly impacts the prediction of classification accuracy.

4. Discussions and Conclusion:

In terms of optimizing CoT-Transfer learning, further research can be conducted to improve the mapping method and the transfer learning network. Currently, the method has a limitation on the length of the sequence, which can be addressed by developing a more effective mapping algorithm. Moreover, the transfer learning network can be enhanced by exploring various architectures that can capture the structural patterns in RNA more efficiently. These efforts can lead to a more accurate and robust method for RNA contact prediction.

Another area for future research is to explore the cuts in RNA in greater depth. While the current definition of cuts is a useful starting point, it may not be appropriate for all RNA sequences. In theory, the more the number of cuts, the more complex the structure of RNA. However, there shouldn't be too many cuts for RNA. Therefore, further studies are needed to investigate the structure of gene sequences to determine the most appropriate definition for cuts. By doing so, we can improve the accuracy of RNA contact prediction and expand our understanding of RNA structure.

Furthermore, in terms of the GCN approach, there are opportunities to optimize the model's structure and improve its explainability. For instance, the input for the GCN is encoded nucleotides, and applying another network to extract a "feature map" for the sequence may enhance the model's performance. Additionally, finding the important node from the GCN is a new challenge, and gradient-based contrast or class activation mapping can be applied to address this issue. These efforts can lead to more accurate and interpretable predictions and enable us to gain new insights into the relationship between RNA structure and function.

Overall, the use of contact matrices has shown great promise in advancing our understanding of RNA structure and function. By addressing the limitations of the current methods and exploring new avenues of research, we can further improve the accuracy and efficiency of RNA contact prediction and expand our understanding of the complex relationship between RNA structure and function. Additionally, the approach can be extended to other areas of bioinformatics and genomics research, such as protein structure prediction and drug design, opening up new possibilities for breakthroughs in the field of molecular biology.

5. References:

Jian et al (Forthcoming), Knowledge from Large-Scale Protein Contact Prediction Models can be Transferred to the Data-Scarce RNA Contact Prediction Task, *Nature Machine Intelligence (submitted)*

Rahnavard, A., Baghbanzadeh, M., Dawson, T., Sayoldin, B., Oakley, T., & Crandall, K. (2023). deepBreaks: a machine learning tool for identifying and prioritizing genotype-phenotype associations.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907.*

6. Appendices:

Appendix A: Results of cut

Saved in the file named [Cuts Results]

This contains all the boxplot and histogram for cuts in the testset.

Appendix B: Distribution for cuts in each contact pairs

Saved in the folder named [Cuts_output]

This contains the results of each contact pair cut by other pairs, which shows all contacts, their label (correct or incorrect), the ground truth for the data in L * L * 37 format, prediction L * L * 37 for the real probability after the softmax layer, their probability from the prediction (P < 10A), and a list of crossing contacts.