THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

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

Complete Capstone Project by: Aihan Liu, Hsuehyi Lu

Master in Data Science, May 2023

Abstract

The rapid increase in the number of proteins in sequence databases has opened up more possibilities for machine learning-based structure and function prediction. However, the scarcity of data in RNA databases hinders the study of RNA.

The RNA pairwise structural proximity (**contact**) was predicted through transfer learning using a protein-coevolution-Transformer-based model. Based on the result of contact matrix, the predicted results and RNA structure were further examined by studying the number and distribution of intersecting contacts (**cuts**).

The functional part of our approach involves using the deepBreaks and functions of sequences to explore the feature importance and provide explanations for the results from co-transformer model. A graph convolutional networks (GCN) is also implemented to learn structure-function relationships.

Our results demonstrate the effectiveness of the method. It achieves high accuracy in contact prediction and provides insights into the important position in the sequence. Our GCN model further enhances the accuracy of our predictions by capturing complex interactions within the RNA structure.

Concepts & Background

- **RNA:** RNA utilizes four bases (A, U, C, G). RNA sequence is a combination for these four nucleotides.
- **RNA contact:** tertiary nucleotide-nucleotide interactions are critical in determining RNA structure and function. If the physical distance is less than 10Å (angstrom, 1Å= 0.1nm) will be considered as "contact".
- **CoT-Transfer Learning:** transfer learning from a protein-based model can improve RNA contact prediction, reducing the data scarcity bottleneck for RNA structural prediction. This finding opens new avenues for research in transferring structural patterns from proteins to RNAs.
- **deepBreaks:** a tool that uses machine learning algorithms to identify and prioritize important positions in genotype-phenotype associations by fitting multiple models and selecting the best one based on cross-validation score, and then using that model to predict the phenotype based on the provided sequence and interpreting it to find the most discriminative positions.
- **GCN:** a neural network designed to work with graph-structured data. They use a convolutional operation to aggregate information from a node's local neighborhood, enabling effective learning and prediction tasks on graphs.

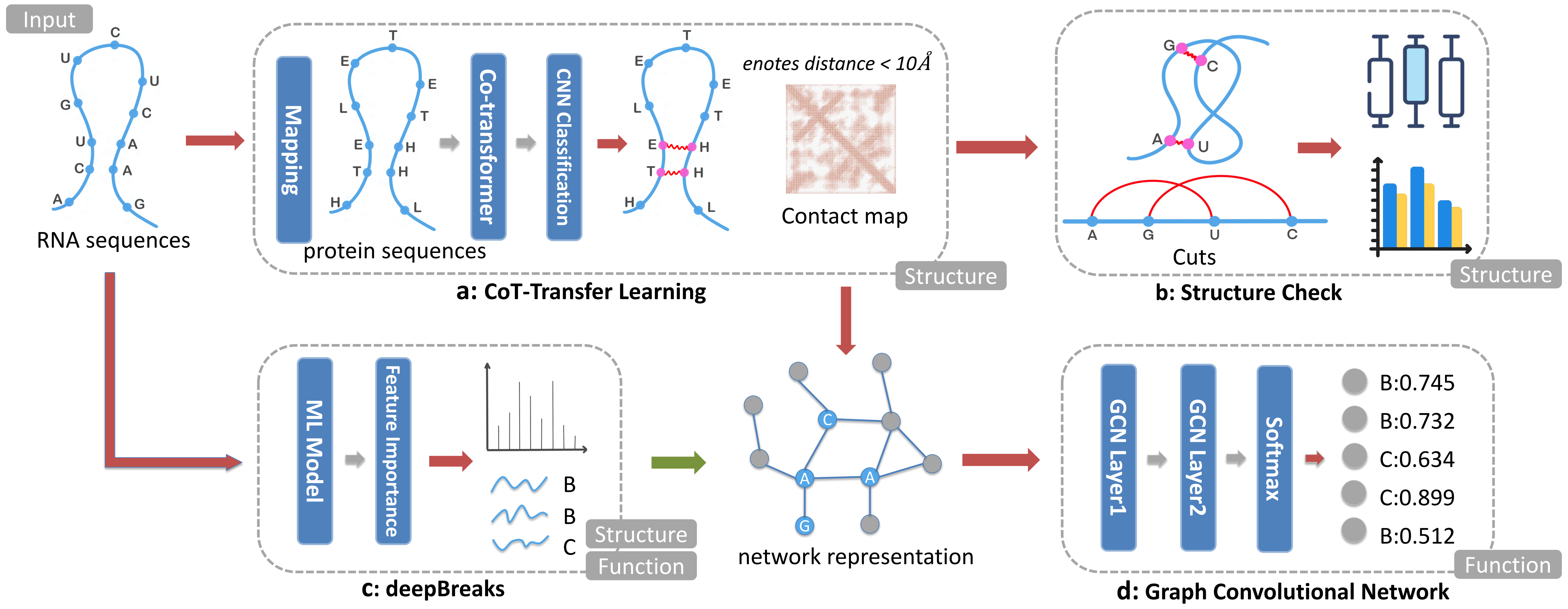
References

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.*

Methodology



The input is RNA sequences.
For the function learning tasks, additional inputs are the label for each sequence.

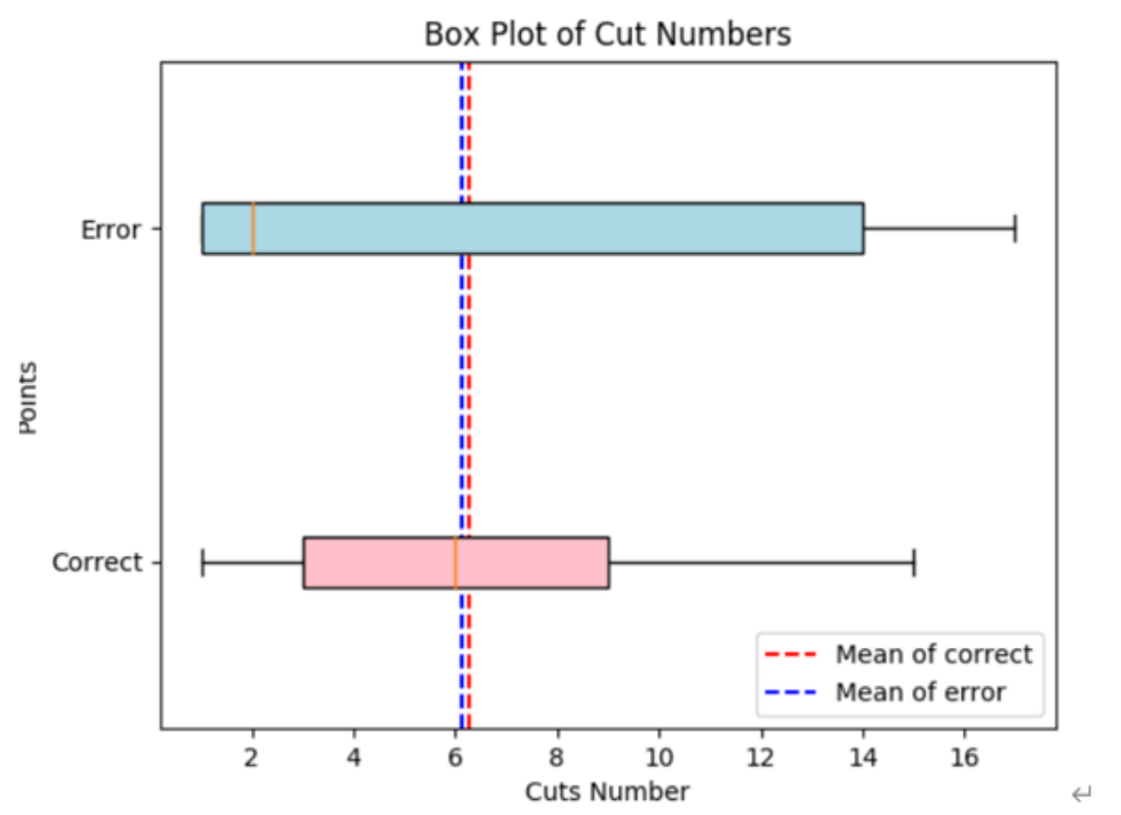
a: A three-stage method to predict the RNA contact based on the protein co-transformer model.

b: Study of number and distribution for cuts from the prediction based on Co-Transferred result.

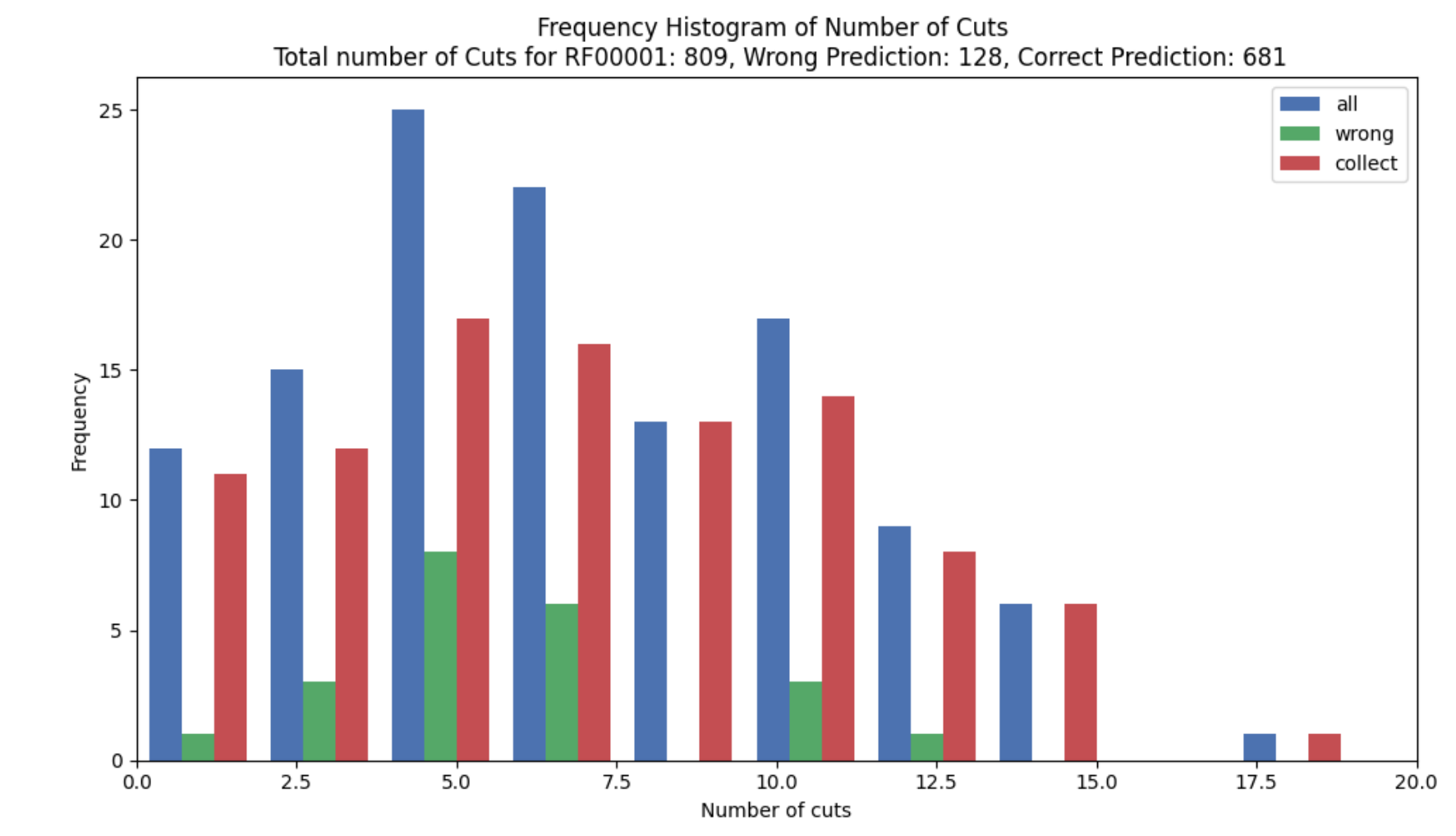
c: Used the deepBreaks to predict the function for RNA. The most important features could overlap with contact map, which provide an explainable result for RNA structure.

d: A GCN with two graph convolutional layers for learning complex structure–function relationships.

Results

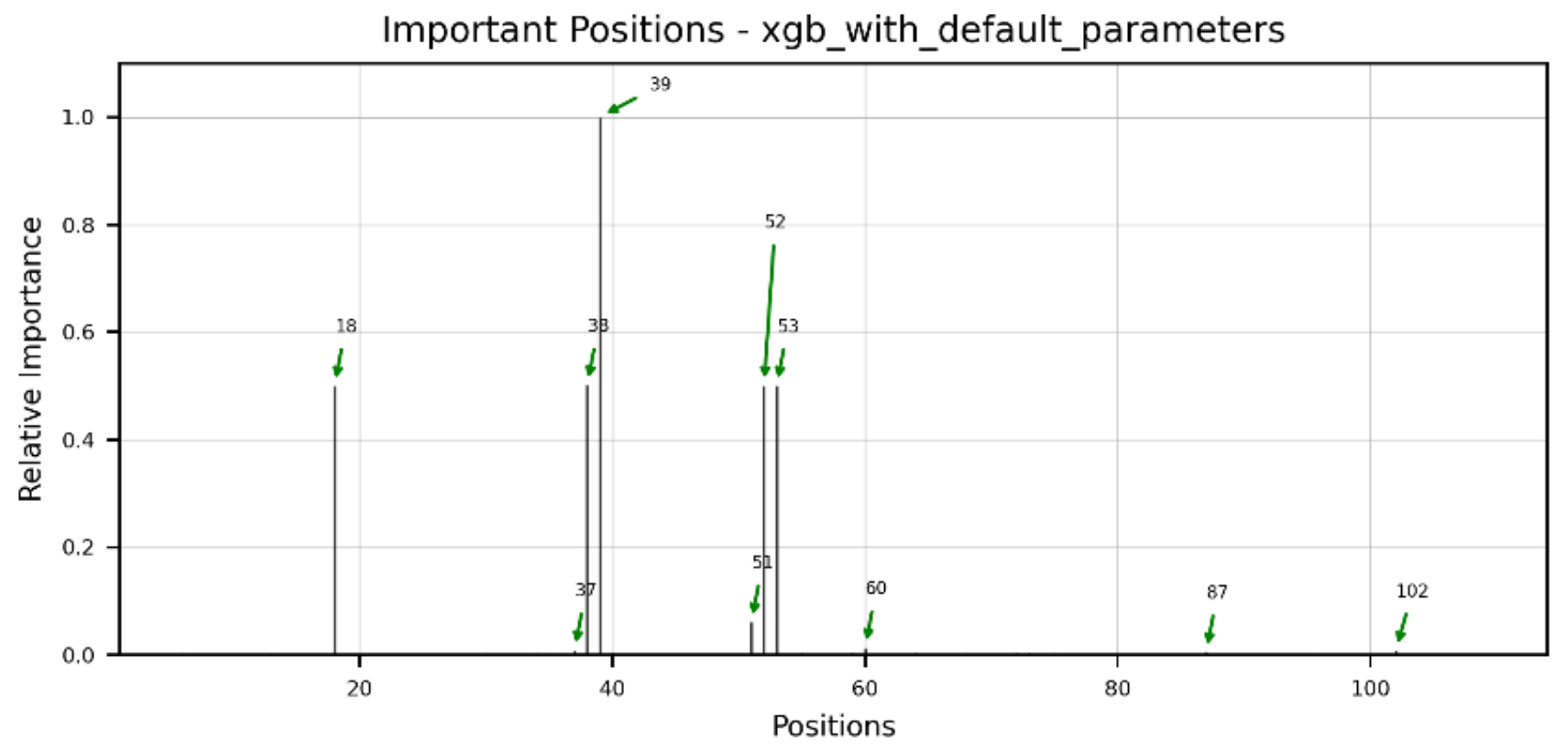


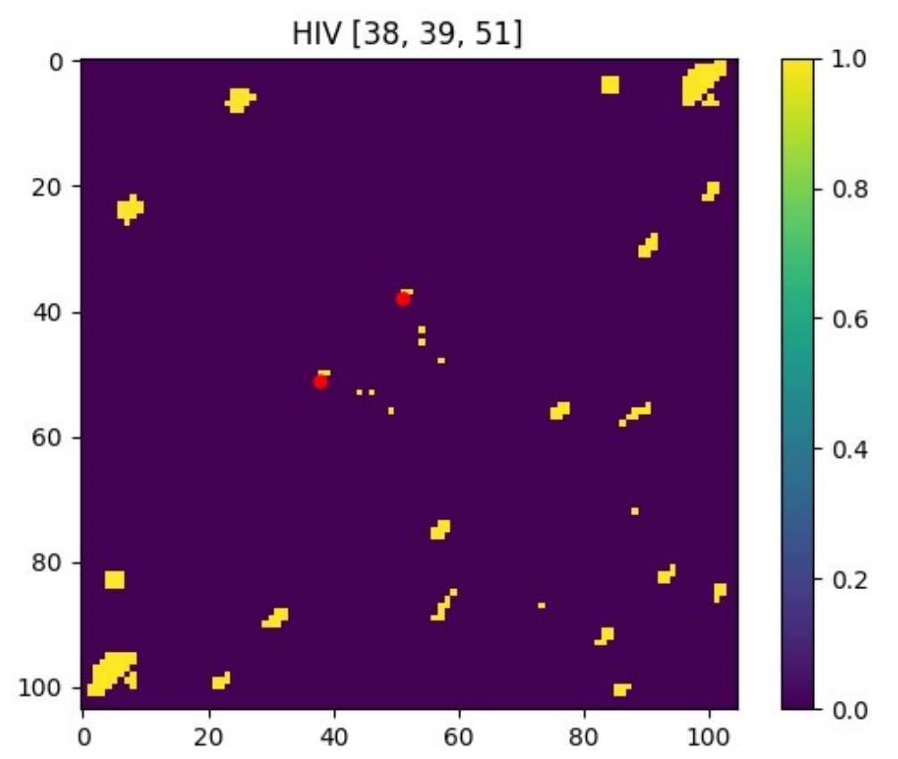
Avg for error:6.09 **Avg for correct: 6.27**



There is no significant pattern for the number and distribution of cuts. It couldn't help us interpret the prediction. Further study is required.

Certain positions appear significantly more relevant than others. Some of these points match the contact matrix, as shown by the red points in the figure.





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

With large samples for training, the contact doesn't help increase prediction accuracy, and the type of nucleotides and order are more critical.

While if the number of samples is smaller, the contact significantly impacts the prediction of classification accuracy.

Future Direction

- **Optimize CoT-Transfer learning:** including improving the mapping method and the transfer learning network. The method has a limitation on the length of sequence.
- **Further study of cuts:** our definition for cuts may not be appropriate. Theoretically, the more the number of cuts, the more complex the structure of RNA. There shouldn't have too many cuts for RNA. It needs more research on the structure of the gene sequences.
- **GCN structure:** the input for GCN is encoded nucleotides. If we apply another network to extract a "feature map" for the sequence, it may have better results for prediction.
- **GCN explainability:** finding the important node from GCN is a new challenge. Some method that could be applied, such as gradient-based contrast, class activation mapping.

In addition, it may also be possible to apply this approach to other areas of bioinformatics and genomics research, such as protein structure prediction and drug design. Overall, the use of contact matrices shows great promise in advancing our understanding of RNA structure and function and has the potential to lead to new discoveries and breakthroughs in the field of molecular biology.

Acknowledgements

We would like to thank our supervisor Professor Chen Zeng for his guidance and support throughout this project.

We would like to express our appreciation to Professor Edwin Lo for his assistance in providing valuable insights into this research.

Special thanks to our classmates for listening and providing suggestion into our research.