Sequence Is All You Need For Accurate RNA-Distance Prediction

Author Name
Affiliation
pcchair@ijcai-22.org

Abstract

Many biological tasks have been well tackled with data accumulation and machine learning advances, particularly protein-related research. However, RNA structure prediction remains a significant challenge in the field due to RNA data limitations. To offer accurate forecasts of RNA 3D-structure, we propose such a task in defining the distance of arbitrary bases in the RNA primary sequence. This regression task is more informative to subsequent 3D folding methods but more complicated than the well-known RNA secondary structure prediction.

In this work, we reveal that with only primary sequential information, we can gain accurate inferences on RNA bases' distance with a sizeable pre-trained RNA language model and a well-designed downstream transformer followed by a unrolled constraint layer. Our experiments show that we outperform all convolutional-based models by a preferably big gap while obtaining rather good statistical results. Moreover, we also acquired a comparable performance with other methods at the contact forecast level by degrading our distance prediction output. Moreover, our approach unified the view of language modeling and distance regression, a new perspective by viewing each predicted embedding as a column vector of the decomposed distance matrix. Our framework will foreseeably be a good guidance for 3D-structure prediction.

1 Introduction

2 Methodology

The whole framework could be mainly divided into four stages, lan pre-training , DiT pre-training, distance map training, and finally the inference stage.

In order to achieve accurate RNA-Distance predictions from vanilla Sequence

2.1 RNA-FM Pre-Training

The intention of this pre-training stage aims to provide rich RNA sequence representations for further downstream tasks as prescribed. A Bert-based language model with 12 transformer encoder blocks [Devlin *et al.*, 2018] was trained on around 26 million non-coding RNA sequences in an unsupervised manner, where details could be found via another work [*]. After the training stage, a learned embedding layer will map an RNA sequence of Length L to a $L \times 640$ tensor.

Noticed that this trained RNA Bert model could be applied directly for fine-tuning in other tasks. However, the difficulty of such an approach lies in the gap between enormous model capacity and relatively small downstream datasets. Thus, we reimplement a transformer-based downstream model DiT specified for tackling the distance prediction task.

2.2 DiT pre-training

F

2.3 Distance Map Tuning

DiT pre-training enbales

3 Methodology

4 Results

5 Results

6 Acknowledgement

7 Acknowledgement

The *IJCAI–22 Proceedings* will asdasdsads

References

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

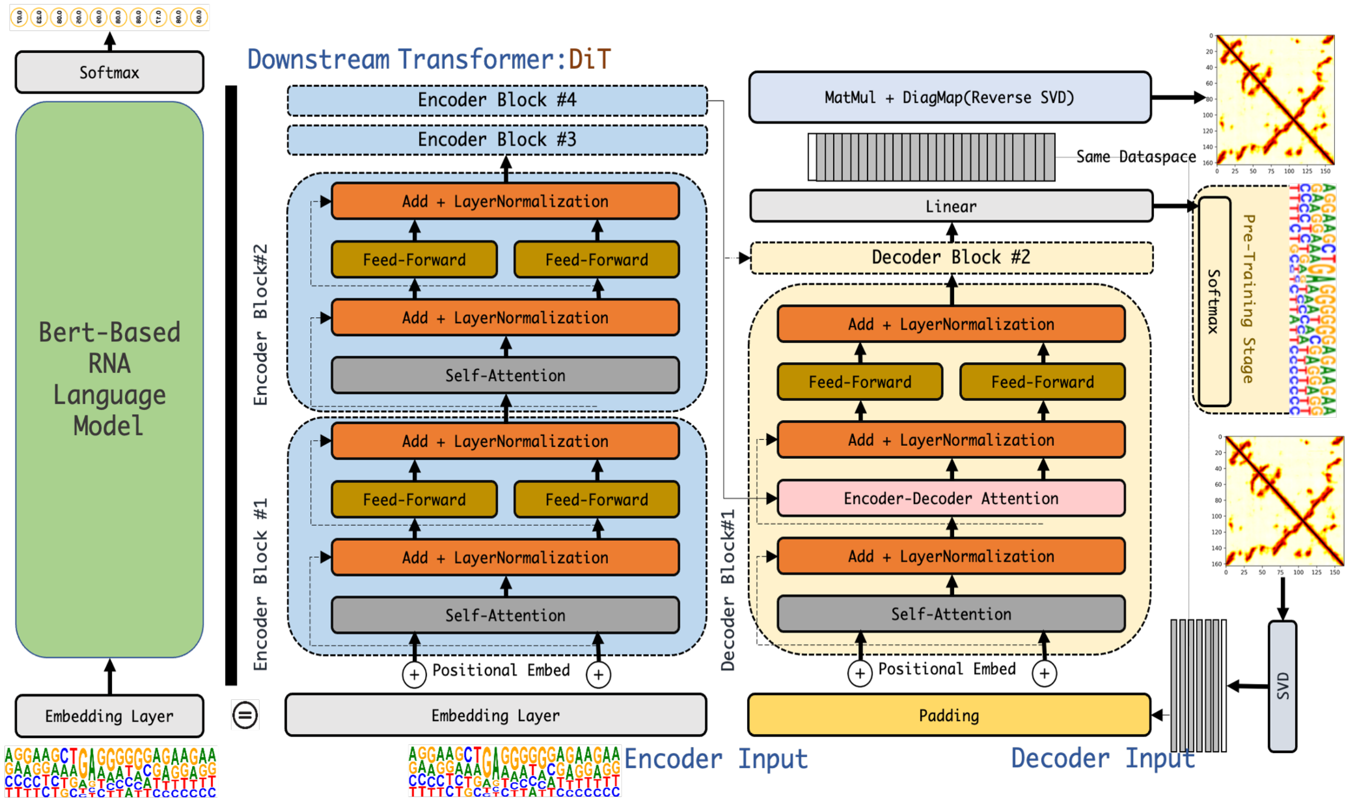

Figure 1: Overview of the model's traininaag Stage

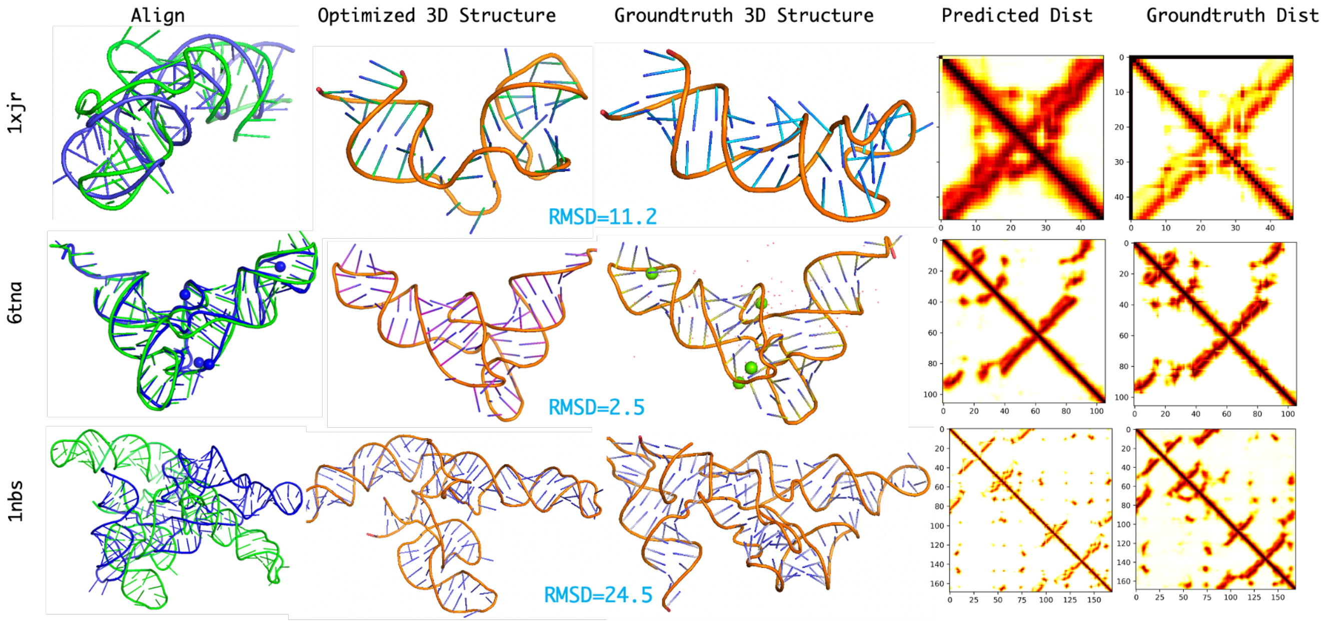

Figure 2: Overview of the model'as training Stage

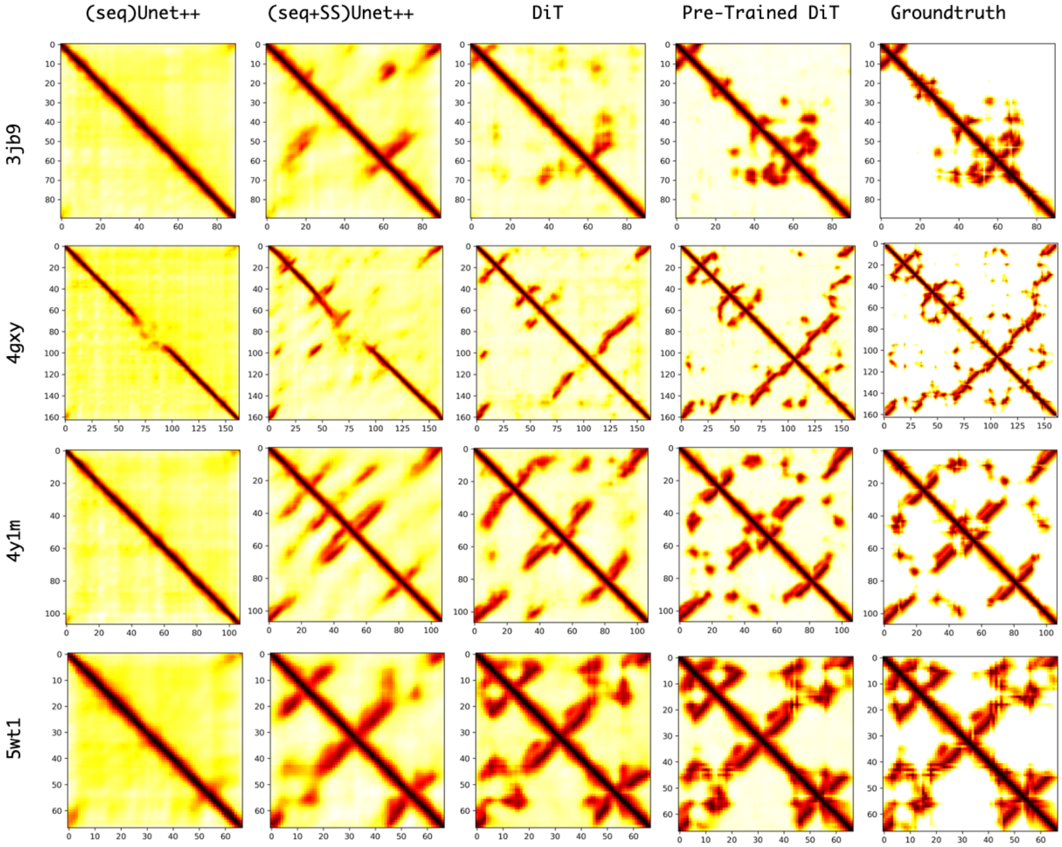

Figure 3: Overview of the modelas's training Stage

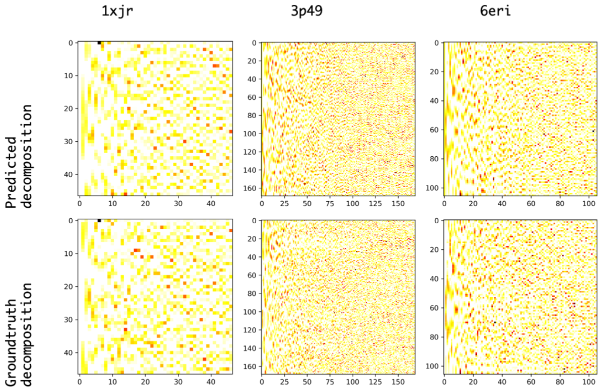

Figure 4: Overview of the modelas's training Stage

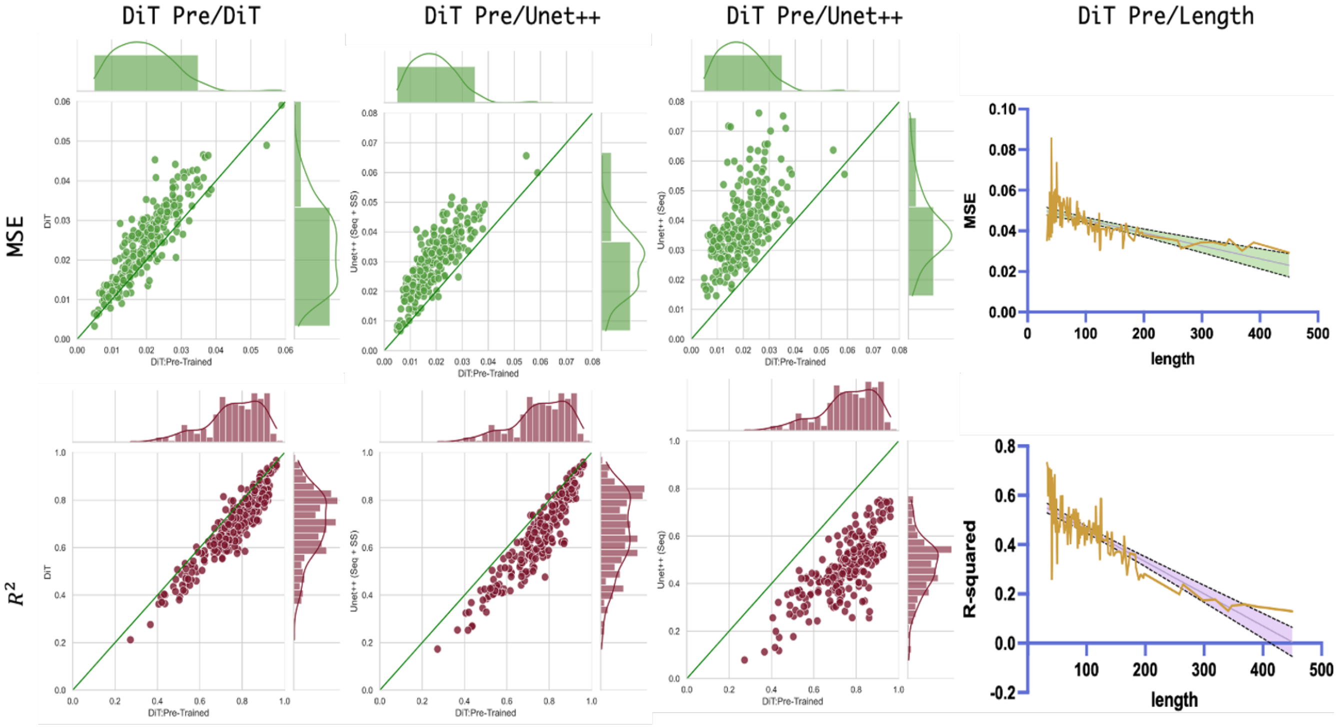

Figure 5: Overview of the model's traiaaaaning Stage