Beyond Sequence: Impact of Geometric Context for RNA Property Prediction

Junjie Xu [1,2 †], Artem Moskalev [1 †], Tommaso Mansi [1], Mangal Prakash [1 ‡], Rui Liao [1 ‡]

[1] Johnson & Johnson Innovative Medicine, [2] The Pennsylvania State University, [†] Shared first, [‡]Shared last

Datasets

	Tc-Ribo	Ribonanza	COVID	Fungal
Dataset Size	Small	Medium	Medium	Large
Task Level	RNA-level	Nucleotide-level	Nucleotide-level	RNA-level
Target	Switching Factor	Degradation	Degradation	Expression
# Sequences	355	2260	4082	7089
Sequence Length	66 - 75	177	107 - 130	150 - 3063
# Labels	1	2	3	1
# Avg. Atoms	1531	3791	2598	N/A

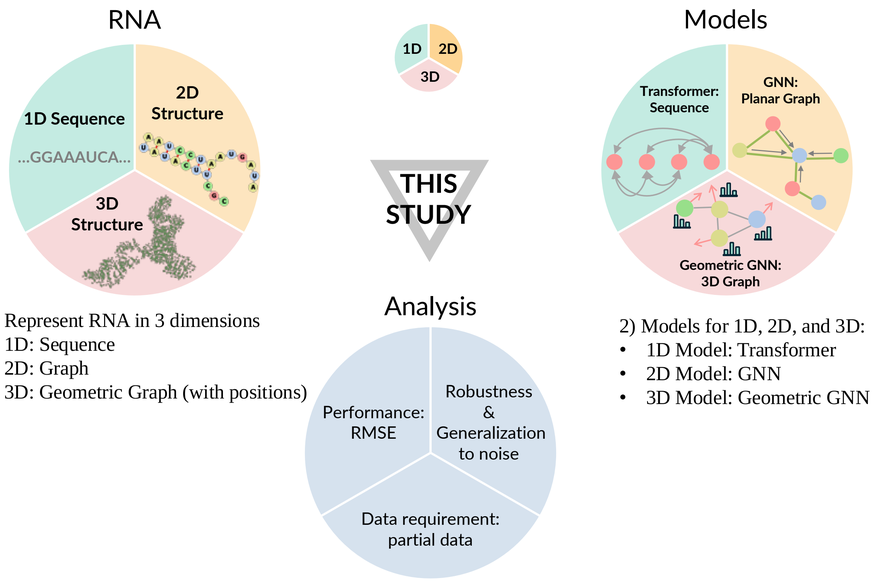
Impact of structural information

Performance: 2D>3D>1D

Model	COVID	Ribonanza	Tc-Ribo	Fungal
1D model				
Transformer1D	0.361±0.017	0.705±0.015	0.705±0.079	1.417±0.005
RNA-FM	0.591±0.081	0.990±0.144	0.693±0.001	1.420±0.028
SpliceBERT	0.588±0.077	1.022±0.144	0.708±0.003	1.435±0.059
2D model				
Transformer1D2D	0.305±0.012	0.514±0.004	0.633±0.001	OOM
GCN	0.359±0.009	0.595±0.006	0.701±0.004	1.192±0.077
GAT	0.315±0.006	0.534±0.006	0.685±0.024	1.112±0.035
ChebNet	**0.279±0.007**	**0.468±0.002**	**0.621±0.022**	**0.973±0.003**
Graph Transformer	0.318±0.008	0.515±0.001	0.710±0.041	1.317±0.002
GraphGPS	0.332±0.013	0.523±0.003	0.715±0.012	1.025±0.081
3D model (without pooling)				
EGNN	0.480±0.025	0.808±0.023	0.725±0.002	OOM
SchNet	0.499±0.003	0.843±0.004	0.696±0.008	OOM
FAENet	0.486±0.010	0.834±0.003	0.703±0.011	OOM
DimeNet	0.497±0.012	0.855±0.006	0.712±0.004	OOM
GVP	0.467±0.010	0.797±0.012	0.744±0.004	OOM
FastEGNN	0.477±0.005	0.816±0.014	0.753±0.001	OOM
3D model (with nucleotide pooling)				
EGNN (pooling)	0.364±0.003	0.619±0.007	0.663±0.010	OOM
SchNet (pooling)	0.390±0.006	0.685±0.006	0.655±0.038	OOM
FastEGNN (pooling)	0.444±0.003	0.753±0.015	0.710±0.011	OOM

Overview

New RNA datasets with 1D, 2D, and 3D structures enable diverse tasks simulating real-world scenarios and evaluating 1D, 2D, and 3D baselines.



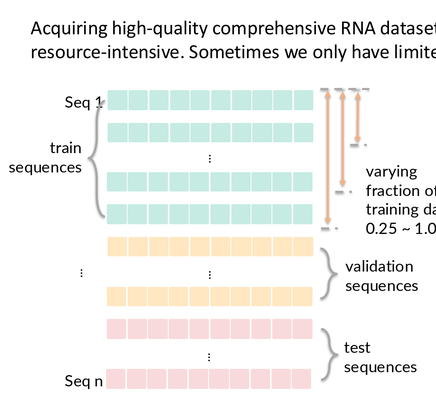
1) Represent RNA in 3 dimensions
- 1D: Sequence
- 2D: Graph
- 3D: Geometric Graph (with positions)

2) Models for 1D, 2D, and 3D:
- 1D Model: Transformer
- 2D Model: GNN
- 3D Model: Geometric GNN

3) Different tasks for different real-world scenarios
- impact of structural information in clean and ideal settings
- The process of acquiring RNA may have noises and errors
- data is limited

Model Efficiency in Limited Training Data Settings

Acquiring high-quality comprehensive RNA datasets is often challenging and resource-intensive. Sometimes we only have limited labeled data for training



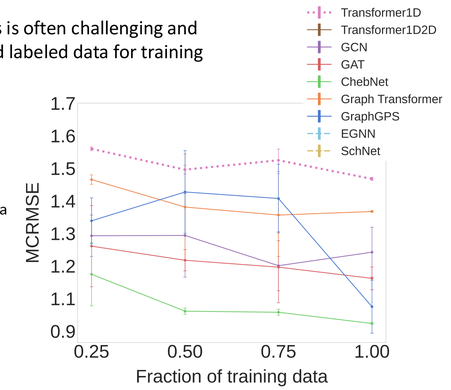
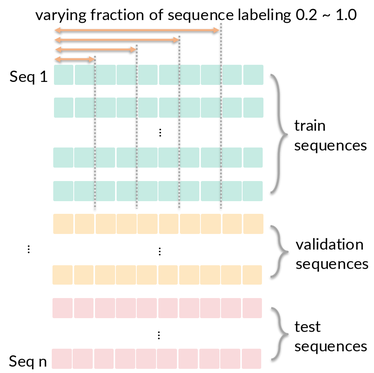


Fig 1. Performance vs. fraction of training data on COVID

- 2D models excel in low data regimes
- 3D models outperform 1D model

Model Efficiency in Limited Training Data Settings

Due to the high cost of measuring properties for every nucleotide in RNA sequence, real-world datasets often contain partial annotations where labels are only available for the first small part of the sequence.



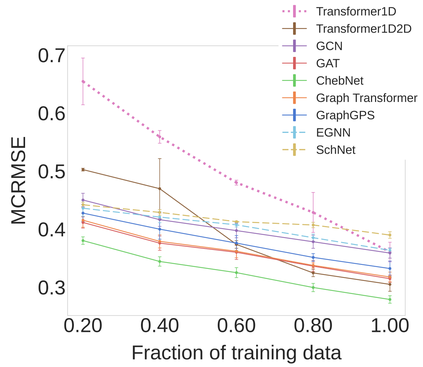
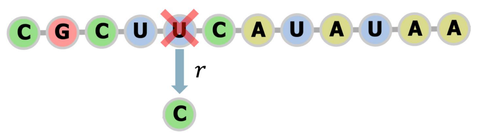


Fig 2. Performance vs. partial property labels on COVID.

- 2D models excel in partial label regimes
- 3D models outperform 1D model

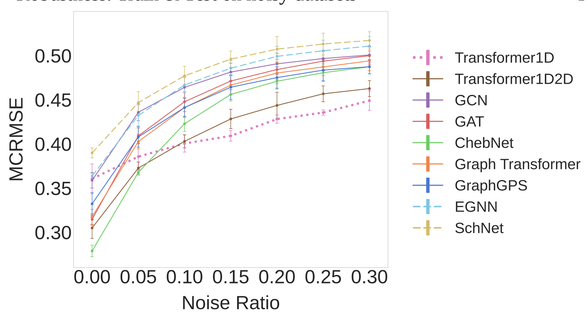
Noisy Data: Generalization & Robustness

RNA is particularly susceptible to sequencing noise due to variability in platforms and quality.



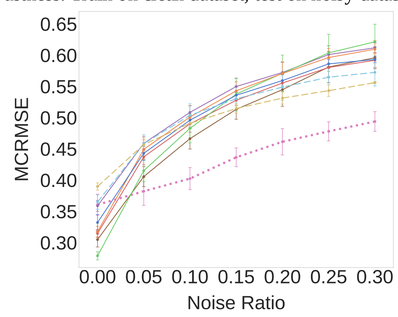
Mutate the nucleotide at each position independently with the probability r.

Robustness: Train & Test on noisy datasets



- Transformer1D shows the least performance drop under increasing noise, maintaining the highest accuracy.
- In contrast, 2D and 3D models, particularly ChebNet and 3D models, are more impacted by noise.

Robustness: Train on clean dataset, test on noisy datasets



- Transformer1D outperforms other models on noisy sequences, achieving the lowest RMSE at higher noise levels.
- Transformer1D2D follows closely, showing that transformer-based models generalize better under noise than 2D and 3D models, especially in tasks with geometric representations.

Contact:
junjiexu@psu.edu
Amoskal2@its.jnj.com
MPraka12@its.jnj.com

Paper