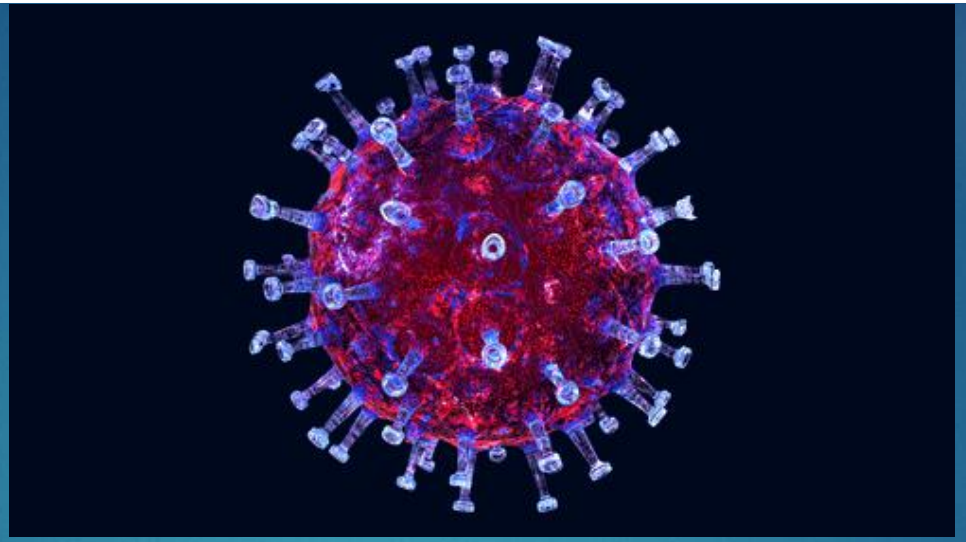
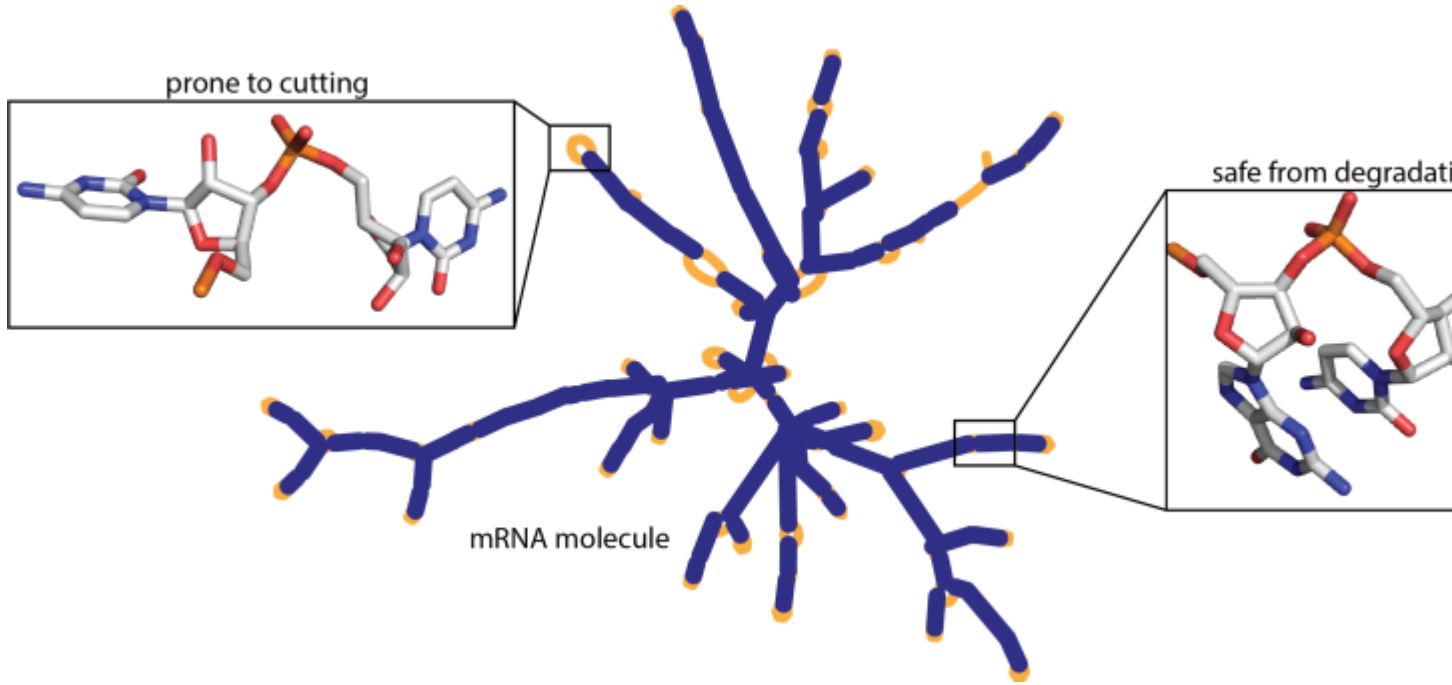


COVID-19 mRNA Vaccine Degradation Prediction

PRESENTED BY: POA MIAO ER

Content

- Background & Problem Statement
- Dataset
- Exploratory Data Analysis (EDA)
- Model Selection
- Model Pre-processing
- Modelling
- Evaluation
- Conclusion & Recommendation



Background

- mRNA (messenger RNA) carry a string of information from DNA. This message will instruct our body to produce the corresponding protein.
- Advantages: easy development process, low time required to develop, reduced risk of pre-existing immunity against the vaccine and a less expensive process.
- Disadvantage: mRNA molecules tend to degrade quickly. Therefore, vaccine are kept under intense refrigeration.

Problem Statement

- To help researchers to understand how are the reactivity and degradation rates of mRNA at different conditions.
- To design model which can predict the reactivity and degradation rates at each base (A, C, G and U) of mRNA molecule.
- To accelerate mRNA vaccine research and deliver a refrigerator-stable vaccine against COVID-19.

Dataset

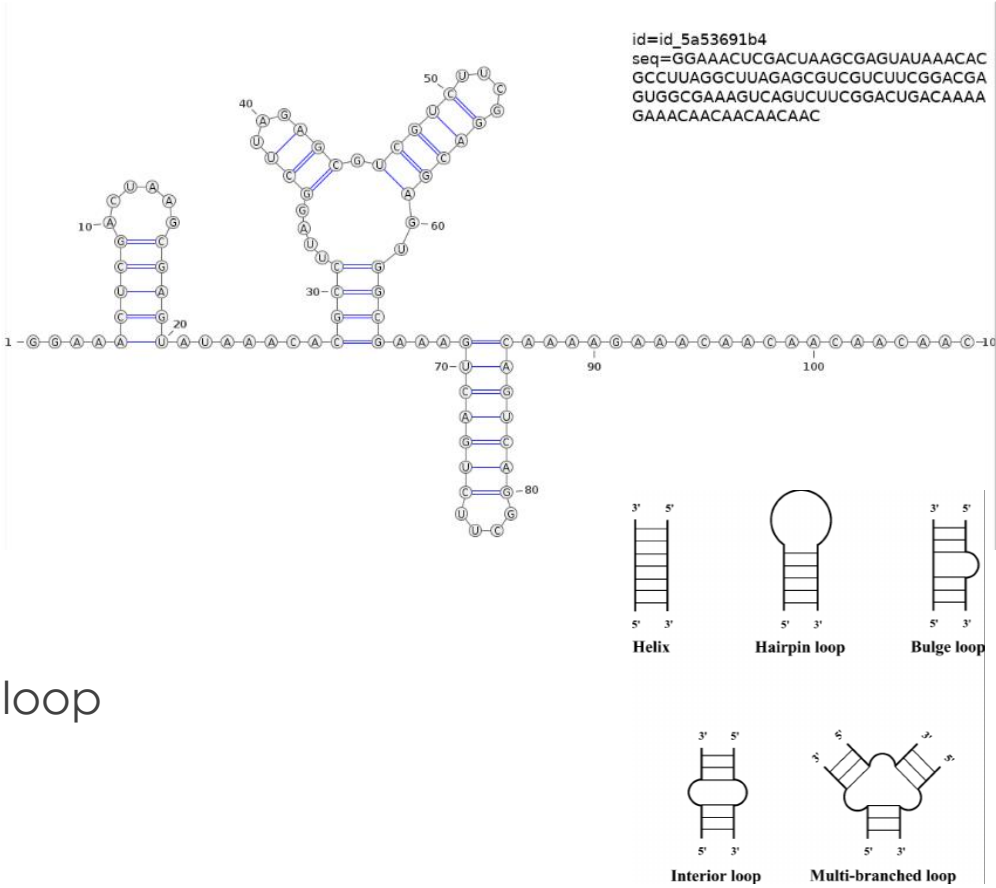
- Data source: Kaggle competition OpenVaccine
- Training Data: 17 columns, 2400 rows
- Testing Data: 5 columns, 3634 rows
- The feature fields (X) are:
  - sequence: describe the RNA sequence
  - structure: whether a base is paired (‘(‘,’)’) or unpaired (‘.’)

    eg. ....((((.....))))........((((.....
  - predicted_loop_type: the structure context

    eg. EEEESSSSHHHHHHSSSS

    (S: Stem M: Multiloop I: Internal loop B: Bulge H: Hairpin loop

    E: dangling End X: external loop)



Dataset (continued)

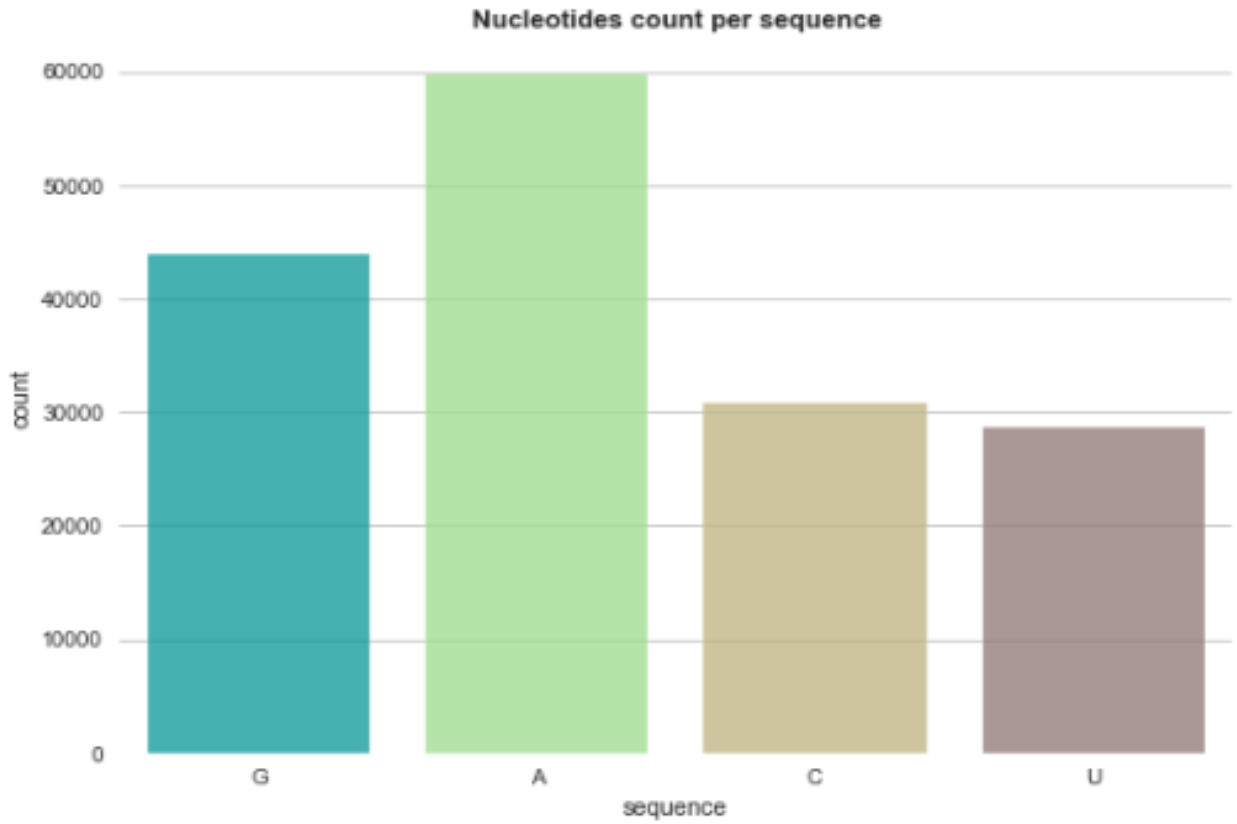
- The predicted fields(y) are:

  - reactivity: determine the likely secondary structure of RNA

  - deg_Mg_pH10 and deg_pH10: determine the probability of degradation at the base after incubating with/without Mg at pH10

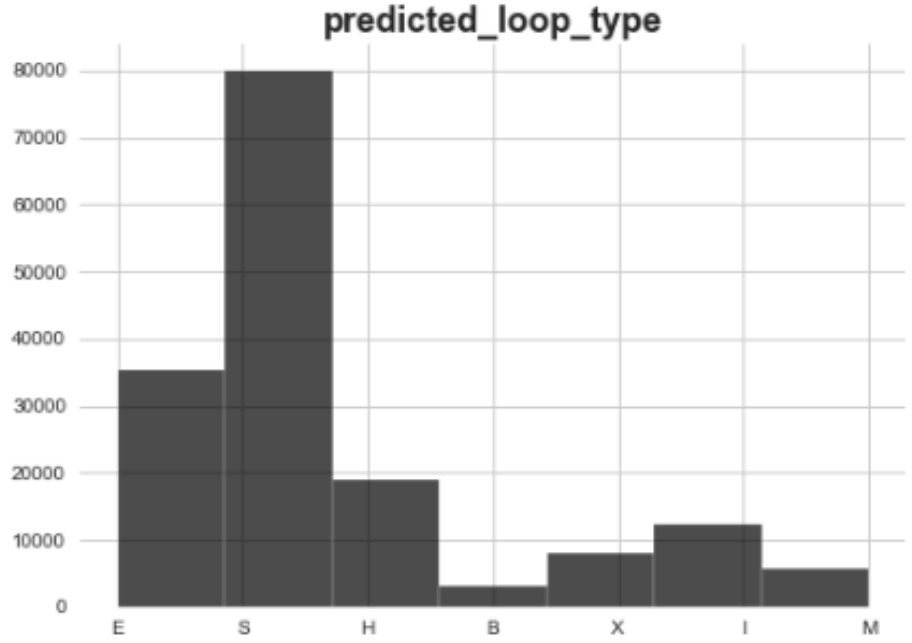
  - deg_Mg_50C and deg_50C: determine the probability of degradation at the base after incubating with/without Mg at 50C

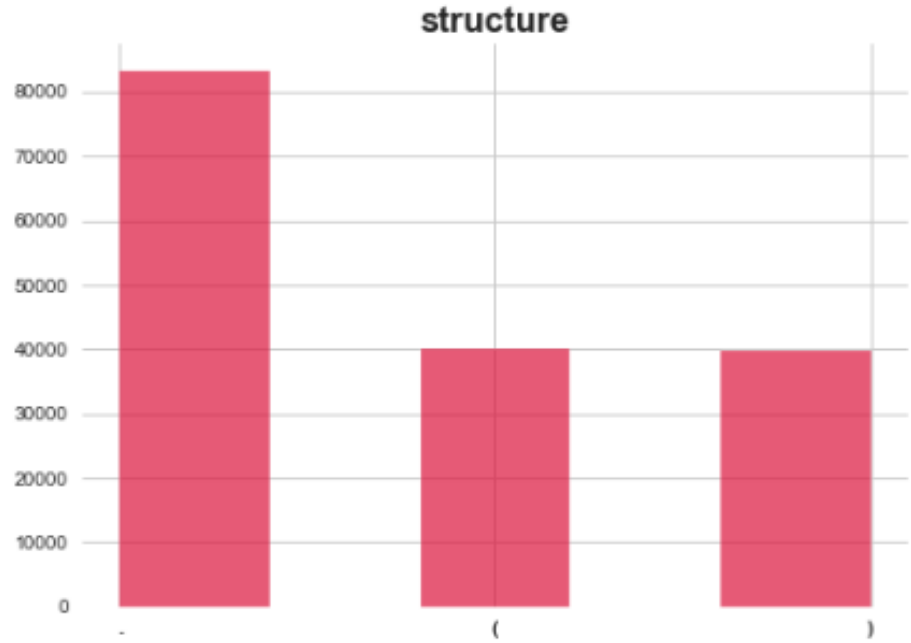
Exploratory Data Analysis (EDA)



- A and G nucleotides are highly present in the sequences compared to C and U.

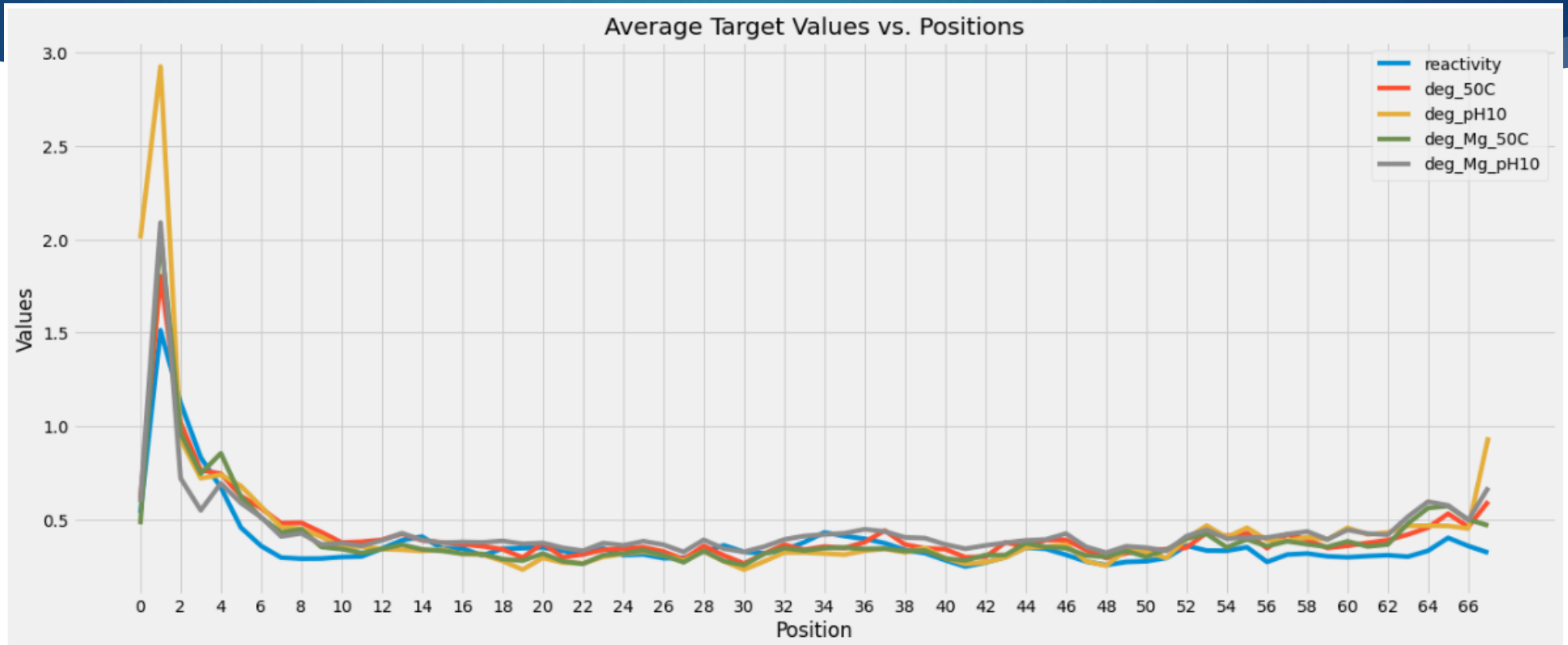
EDA (Continued)





- S (Stem = paired) is the dominant loop type.
- E (Dangling End = unpaired terminal nucleotides) and H (Hairpin Loop) are also highly represented in comparison with the rest.
- . structure (unpaired) is dominating, the paired structures) and (are equally represented (since their pair together).

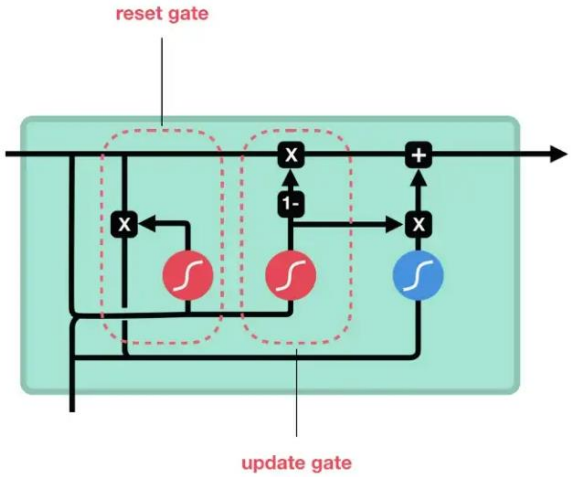
EDA (continued)



- High values of Degradation & Reactivity at the beginning of the sequence.
- There is high degradation because of pH10 at the beginning but no such pattern for rest of the positions.

Model Selection

- mRNA is a sequence of bases and each base is dependent on the bases that come before it.
- RNN suffers from short-term memory and thus often suffer from vanishing gradients problem. One of the solution for this is GRU.
- Gated Recurrent Unit (GRU) has two gates.
  - **Reset gate** determines how to combine new input to previous memory.
  - **Update gate** determines how much of the previous state to keep.



Model Pre-processing

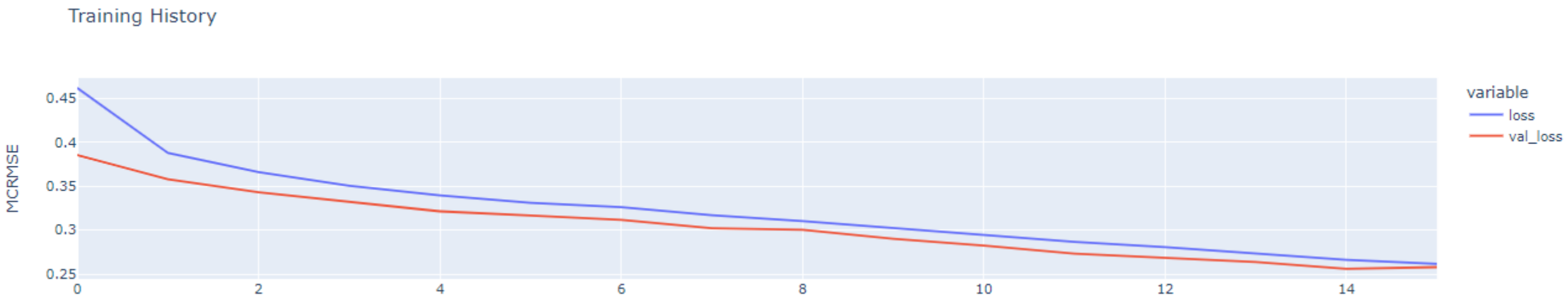
- Train-Test Split: 90% and 10%
- Use stratify to filter the training data with ‘signal_to_noise’ feature to exclude the noisy samples that have values ≤ 1 (outlier)

Modelling

- Embedding layer: The categorical features are encoded by numbers, and then, the features are extracted for learning.
- Hidden layer: Bidirectional GRU layer is used to optimize the results as the data are passed in the forward and backward directions to better capture the information in the sequence data.
- Dense layer: The dense layer has 5 outputs (five target columns of reactivity and degradation) and the activation = ‘linear' because the problem is a regression problem.
- loss = 'MCRMSE’: there are multiple outputs that we are trying to predict. The MCRMSE is simply an average across all RMSE values.

Evaluation

- Train and test loss are similar. Loss curve of both train set and test set follow each other closely and decreases overtime. Hence our model is not overfit.
- Loss: 0.2614 vs Val_loss: 0.2578



Conclusion & Recommendation

- GRU model is adopted for model deployment.
- Finding: high values of degradation & reactivity at the beginning of the sequence.
- By combining the finding with the prediction model that we developed, we are able to help researchers in designing more stable covid-19 mRNA vaccine.
- Future work: Tune parameters, Stack bidirectional GRU with bidirectional LSTM, try CNN.

Model Deployment