COVID-19 mRNA VACCINE DEGRADATION PREDICTION

Enrolment. Nos. : 17103314, 17103359

Name of Students : Sankalp Biswal, Shobhit Agarwal

Name of Supervisor : Dr. Dhanalekshmi G, Dr. Adwitiya Sinha

May-2021

Major 2 (Final Report)

Submitted in partial fulfilment of the Degree of

Bachelor of Technology
In
Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

DECLARATION

We hereby declare that this submission is my/our original work and that, to the best of my knowledge and belief, it contains no material that has been previously published or written by another individual, nor does it contain material that has been approved for the award of any other degree or diploma by the university or other institution of higher learning, except where appropriate acknowledgement has been made in the document..

Date: 20th May 2021

Name(s): Sankalp Biswal, Shobhit Agarwal; 17103314, 17103359.

CERTIFICATE

This is to certify that the work titled **COVID-19 mRNA VACCINE DEGRADATION PREDICTION** submitted by Sankalp Biswal, Shobhit Agarwal in partial fulfilment for the award of degree of B.tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or entirety to any other University or Institute for the purpose of awarding this or any other degree or diploma.

Name of Supervisors : Dr. Dhanalekshmi G, Dr. Adwitiya Sinha
Date: 20th May 2021

Acknowledgement

This project is an outcome of a long pending interest and curiosity in the clustering and topic modelling techniques. We have invested a great deal of time and energy in this endeavour. It would not, however, have been possible without the generous support and assistance of numerous individuals and the institute. As such, we would like to express our heartfelt gratitude to each of them.

We are extremely appreciative of our supervisor, Dr. Dhanalekshmi G, and Dr. Adwitiya Sinha, for their guidance, continuous supervision, and for providing necessary information about the project and holding meetings as required.

Additionally, we would like to express our gratitude to our parents and institute faculty members for their kind cooperation and encouragement, which aided in the completion of this project. Our gratitude and appreciation also go to our friends involved in the project's development and to others who have volunteered their time and abilities to assist us.

Name of Students: Sankalp Biswal, Shobhit Agarwal

Enrolment Number(s): 17103314, 17103359

Date : 20th May 2021

Table of Contents

1. INTRODUCTION

1.1. General Introduction

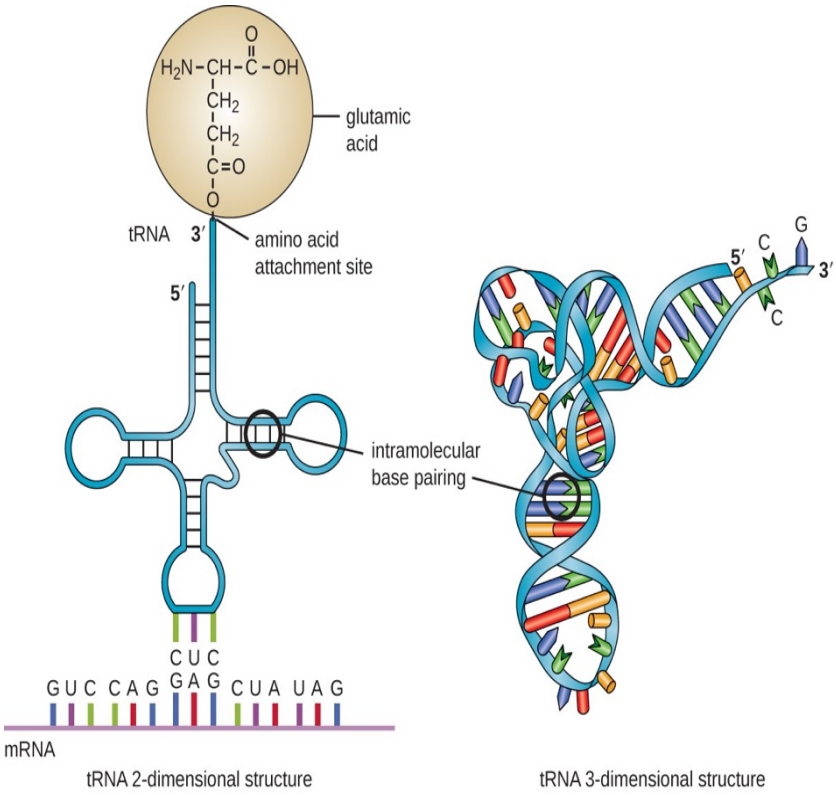
To end this pandemic, it will be critical to have an effective COVID-19 vaccine, which will be disseminated to everyone on a reasonable and equitable basis. Scientists have built on decades of the previous study to advance the search for a vaccine against COVID-19, and if one more day were lost, billions of dollars in damage to the ecosystem would occur. We are interested in receiving new, creative ideas from all across the world. Researchers and crowdsourcing knowledge could lead to measurable improvement in computational biochemistry. Despite being the quickest candidates currently in the race, the candidates for COVID-19 may have to yield to mRNA vaccines in the future. Making messenger RNA molecules that are exceedingly stable is a daunting task right away. As of the time, mRNA vaccines cannot be carried by standard means, as only disposable syringes are available and these are shipped with chilled shipments around the world that are refrigerated.

RNA molecules will die over time due to spontaneous decay. Cutting just one mRNA in the vaccination renders it useless. Currently, the location of the RNA's most sensitive parts is unclear. To produce the mRNA against COVID-19, this data is needed, which means it is quite unlikely that a vaccine targeting this segment of the population will be an influence on more than a small fraction of the global population. In this process, we utilize our previous knowledge about how to utilize the data science to come up with and enforce model and style requirements for RNA degradation.

Until now, the majority of research has focused on the stability and destruction of RNA molecules using standard statistical approaches and biophysical models. To what extent do RNA molecules experience a spontaneous breakdown, and to what extent do mRNA molecules respond to environmental changes? Because experimentation is the norm for discovering values such as these, experimentation is the only way of determining if such values are valuable.

Our RNA degradation model will allow us to make estimates about the different parameters which affect the rate of degradation(Eterna dataset). It's possible that in the event of a pandemic, mRNA vaccines may have to be enhanced before we can confirm whether or not they are effective. We employ scientific methods to better grasp scientific problems, which allows us to proceed with our mRNA vaccine research and to bring about a long-term SARS-CoV-2 vaccine. Conventional immunizations are more cost-effective and quicker to produce thanks to molecular vaccines. They are

also safer to administer. Transferring RNA molecules becomes problematic since they can break down on their own.



tRNA 2-dimensional structure

tRNA 3-dimensional structure

The genetic information contained in the nucleus circulates across the cell in the form of single-stranded molecules known as messenger RNAs. Most mRNAs can't be translated, yet proteins are made by using the information in these mRNAs. When referring to the nucleotides, the acronyms A, U, G, and C are commonly used. There are four essential elements of the mRNA: U (uracil), A (adenine), C (cytosine) and G (guanine).

1.2. Problem Statement

This project suggests a way to forecast the stability and potential risk of degradation of a RNA sequence under certain circumstances . The project's goal is to estimate the RNA molecule's stability through the "inputs" and "target values" listed in the table below.

Feature	Classification	Description	Sample
Sequence	*Input*	A sequence of 107 letters corresponding to the four bases in the sequence.	A, G, U, U, C, ...
Structure	*Input*	Expected structure of the molecule (length = 107). '(' and ')' refer to a base pair interaction. All '.' in the middle are associated with no BP interactions.	(..()...)()(... ...
Predicted loop type	*Input*	Predicted secondary structure of the RNA molecule at different points. 'S' refers to a stem structure, 'M' multiloop, 'I' internal loop, 'B' bulge, 'H' hairpin loop, 'E' dangling end, 'X' external loop.	S, S, M, S, H, ...
Reactivity	*Target*	Reactivity values at each individual point in thesequence.	1.23, 3.46, ...
Deg pH 10	*Target*	Degradation values at pH 10.	0.89, 2.44, ...
Deg pH 10 Mg	*Target*	Degradation values at pH 10 with added Mg.	1.28, 0.88, ...
Deg 50°C	*Target*	Degradation values at 50°C.	2.02, 1.87, ...
Deg 50°C Mg	*Target*	Degradation values at 50°C with added Mg.	1.11, 2.44, ...

1.3. Significance of the problem

Interest in RNA-based technology for the generation of preventive vaccinations has increased during the last two decades. Thanks to their efficacy, capability for quick and low-cost manufacture, and relatively safe administration, they are commonly accepted as viable alternatives to typical procedures. The illnesses under study include SARS-CoV-2, although new vaccines have been authorised for human use yet.

One of the first hurdles that new drugs have to face is that the RNA molecule is very fragile, and these treatments must be either freeze-dried or held at low temperatures to avoid deterioration. Within the current outbreak, vaccinations are the most promising way to control the novel coronavirus as with the use of vaccines we can introduce the antibodies to fight the infiltration in the body. However, given the present limitations on mRNA vaccines, it is doubtful that these vaccines can efficiently reach all of the people to defeat corona virus once and for all.

1.4. Brief Description of the Solution Approach

To help make accurate predictions about degradation rates, in this project we are presenting a system that analyses a portion of the Eterna dataset, containing over 3,000 RNA molecules and their degradation rates at each place.

The overall goal of this project is to illustrate a Deep Learning model with the dataset. A variant of Recurrent Neural Networks (RNNs) which has two variants namely, Long Short Term Memory Networks (LSTMs) and Gated Recurrent Unit Networks (GRUs) is used. The model employs the GRU algorithm to see if machine learning methods can deliver benefits in the prediction of mRNA molecules' reactivity and degradation.

2. LITERARY SURVEY LEARNINGS

In this race to find a safe and efficient vaccine against coronavirus illness (COVID-19), pharmaceutical formulation science is critical for the processing, delivery, and immunization procedures. The many vaccine formulations, carriers, vectors, adjuvants, excipients, dose forms, and administration routes all influence the immune responses and the efficiency of COVID-19. In this study, we used Google search engine and search results databases to identify the COVID-19 vaccines that are now being evaluated in clinical trials, and we then went into great detail into the different types of vaccines, their formulations, benefits, and probable limits. Another issue that we explored was how to address vaccine delivery and administration concerns. We figured out how to minimize the potential of these issues through the use of a vaccine-stabilization technique and developed specific ways of mucosal immune response-inducing, non-invasive administration methods, which must be considered early in the development process.

Live attenuated and inactivated pathogens and subunit vaccines are traditional vaccine technologies, such as these, that provide long-term protection against a wide range of hazardous diseases. Although these advancements have been made, the task of developing a broad spectrum of vaccines to fight against infectious diseases, especially those that have more successful evasive strategies, remains difficult. For these new virus vaccines, the greatest obstacle is not efficacy, but rather, the requirement for increased manufacture and implementation on a broad scale. In other words, typical vaccination techniques may not be suitable for conditions other than infectious diseases, such as cancer. Developing more powerful and scalable vaccination platforms is essential to achieving this goal.

It was shown in 1992 that injecting mRNA that contains the information for a portion of the vasopressin protein into the hypothalamus created a physiological response in rats. But despite these first discoveries, funding for developing mRNA-based treatments is still lacking because of worries about mRNA instability, the innate immunogenicity of the mRNA, and poor in vivo delivery. Instead, the industry has moved away from drug delivery and toward the use of DNA and protein-based treatments.

Advances in the production of recombinant DNA vaccines are rapidly occurring, with a large amount of preclinical evidence accrued over the previous several years and many human clinical studies beginning. We examine existing mRNA vaccination approaches, describe the most recent trends, discuss barriers, and mention recent achievements. We also look to the future, discussing the challenges mRNA vaccines will face and predictions on their long-term presence. Many of the vaccine-production difficulties of infectious diseases and cancer may be solved with mRNA vaccines.

2.1. Summary of the literature studied

Serial no.	1
Title of the paper	COVID-19 vaccine development and potential nanomaterial path forward
Contributors	Matthew D. Shin , Sourabh Shukla, Young Hun Chung , Veronique Beiss
Year of publication	2020
Journal/publication	Nature Nanotechnology
Summary	The improvements in bio and nanotechnology as well as the recent advancements in advanced nanotechnology (including open reporting and data sharing) have inspired the quicker advancement of new vaccination technologies in the COVID-19 pandemic, which aims to help improve lives during the global pandemic. There are several possible uses for multiple nanomaterials, which provides the potential for scalability, stability, mobility, and the capacity to integrate self-administration platform technologies. In addition, several of the platform technologies covered in this research have the ability to operate in other seasons or new coronavirus strains, using plug-and-play technologies.

Serial no.	2
Title of the paper	The COVID-19 Vaccine Race: Challenges and Opportunities in Vaccine Formulation
Contributors	Jieliang Wang, Ying Peng, Haiyue Xu,Zhengrong Cui,and Robert O. Williams
Year of publication	2020
Journal/publication	AAPS PharmSciTech
Summary	In this review they offered a concise explanation of vaccine types, their formulation, their advantages, and probable constraints in this evaluation of the vaccinations that are presently being tested in clinical trials for COVID-19. They also highlighted how vaccination stabilization techniques and the utilization of early development of specific immune response generating mucosal, non-invasive administration pathways could be able to solve delivery and administration challenges.

Serial no.	3
Title of the paper	mRNA vaccines — a new era in vaccinology
Contributors	Norbert Pardi, Michael J. Hogan, Frederick W. Porter and Drew Weissman
Year of publication	2020
Journal/publication	Nature review Drug Discovery
Summary	MRNA vaccines present a prospective alternative to traditional vaccination techniques, due to their high capacity, fast development capability, and low manufacturing cost and safe administration. Prior to recently, in vivo distribution of mRNA was ineffective, and therefore they have been relegated to laboratory application. Many mRNA platforms for immunization against infectious illnesses and certain malignancies have demonstrated promising outcomes in both animal models and people, having once been confronted with significant difficulties. This review looks into mRNA vaccines in-depth and weighs in on potential future avenues and challenges in this particular field.

Serial no.	4
Title of the paper	Advances in mRNA Vaccines for Infectious Diseases
Contributors	Cuiling Zhang, Giulietta Maruggi, Hu Shan and Junwei Li
Year of publication	2020
Journal/publication	FrontiersIN
Summary	Prophylactic and therapeutic vaccines have in the past two decades mostly been focused on RNA-based technology. In animal models and in people, clinical studies and preclinical studies reveal that mRNA vaccines are safe and long-lasting. As for mRNA vaccines, which might be manufactured swiftly and will be used to construct strong instruments against infectious illnesses, the researchers review recent work here.

Serial no.	5

Title of the paper	Machine Learning Methods Enable Predictive Modelling of Antibody Feature
Contributors	Ickwon Choi, Amy W. Chung, Todd J. Suscovich, Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Robert J. O'Connell, Donald Francis, Merlin L. Robb, Nelson L. Michael, Jerome H. Kim, Galit Alter, Margaret E. Ackerman, Chris Bailey-Kellogg
Year of publication	2015
Journal/publication	Journal PLOS
Summary	It is possible that the adaptive immune response or an infection may result in the generation of specific antibodies that target the pathogen, as well as cells that can promote the adaptive immune response. While anticorps serve to stimulate cell response, it's possible that they also served as a protective mechanism for the RV144 HIV test. Several machine learning algorithms are utilized to group and predict connections between antibody (igg and antigen-specific) and effector functions in a comprehensive dataset acquired from RV144 participants (antibody dependent cellular phagocytosis, cellular cytotoxicity, and cytokine release).

Serial no.	6

Title of the paper	Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network
Contributors	Alex Sherstinsky
Year of publication	2020
Journal/publication	ELSEVIER
Summary	The Vanilla LS TM network's combination with the LSTM system was used to identify additional unique training possibilities, which led to the creation of the most popular LSTM version to date. Students have previously had opportunities to try out RNNs and LSTM networks due to a wide range of tool options and varied educational approaches. In our study, as well as our test and experimental software, our enlarged LS-TM model was implemented using a machine learner, a type of machine learning model which uses our expanded LS-TM model as a reference.

Serial no.	7

Title of the paper	Comparative Study of CNN and RNN for Natural Language Processing
Contributors	Wenpeng Yin,Katharina , Mo Yu, Hinrich Schütze
Year of publication	2017
Journal/publication	arXivLabs
Summary	DNNs have revolutionized the field of natural language translation (NLP). A CNN (convolutional neural network) and a Recurrent Neural Network (RNN) make up the two main kinds of DNN designs (RNN). Since CNNs may reduce the number of position invariant features and RNN units, CNNs should be beneficial for eliminating these properties from the overall model. A lot of NLP practices are modified by the most recent advances in CNN and RNN battling for dominance. This is the first time we've done a comprehensive CNN-RNN comparison using a wide range of NLP jobs to supply guidance for DNN choices.

Serial no.	8

Title of the paper	Bidirectional Recurrent Neural Network-Based Chemical Process Fault Diagnosis
Contributors	Shuyuan Zhang, Kexin Bi, Tong Qiu
Year of publication	2020
Journal/publication	ACS Publications
Summary	Periodic defect detection is vital in industrial chemical processes in order to ensure that they run safely and reliably. Using the vast amounts of data available from the Big Data era, new ways for complicated chemical processes utilizing data-driven failure detection and diagnosis (FDD) methodologies may be developed. This extensive attention is directed at deep learning-based FDD approaches, which use an artificial neural network to obtain feature data from raw data (ANN). Time series data across different types of neural networks works extremely well with recurrent neural networks (RNNs). With regard to extractability of features, typical unidirectional RNNs only continue in the positive direction with a lower defect diagnosis. bidirectional RNNs were utilized to create complex FDD models with RNN cells featuring nuanced functionality. This outstanding BiRNN-dependent FDD performance highlights how successful BiRNN is in chemical fault diagnosis.

Serial no.	9

Title of the paper	Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)
Contributors	Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille
Year of publication	2014
Journal/publication	arXivLabs
Summary	This article presents a multimodal network model for new picture subtitles (m-RNN). It explicitly describes the probabilities of previous words creating a word and a picture. The picture titles sampled from this distribution are produced. The model has two subsystems: a long-term memory neural net, which is made up of several recurrent convolutional layers, and a deep-learning image processing subsystem. In order to create the whole m-RNN model, the two subnetworks communicate in a multimodal layer. It offers four evaluation datasets for us to test how useful our model is. The model goes much beyond the most advanced approaches. M-RNN is furthermore utilized to retrieve pictures or keywords, considerably improving search engine optimization effectiveness.

Serial no.	10

Title of the paper	LSTM vs. GRU vs. Bidirectional RNN for script generation
Contributors	Sanidhya Mangal, Poorva Joshi, Rahul Modak
Year of publication	2019
Journal/publication	arXivLabs
Summary	This research illustrates how various learning models that assist sequential understanding sequence by sequentially executing scripts in order to develop new character talks and new scenes written from a script. Each model was evaluated in depth, including LSTM, GRU, and bidirectional RNN. The models are designed to learn the sequence of the recurring characters from the sequence of the input data. There are an equal amount of characters in each input sequence, which is the letter "n," and the targets are the same, regardless of where the one character that is shifted is placed. The models are trained by iteratively generating and using input and output sequences. Another way to describe the results of an analysis is by utilizing a graph plot and sentences that were created based on a portion of an input string. In these graphics, every model is shown to have improved performance.

3. REQUIREMENT ANALYSIS AND SOLUTIONS APPROACH

3.1. Requirement Analysis

Our project has been programmed using Python and TensorFlow is used for backend in collaboration with a Keras API.

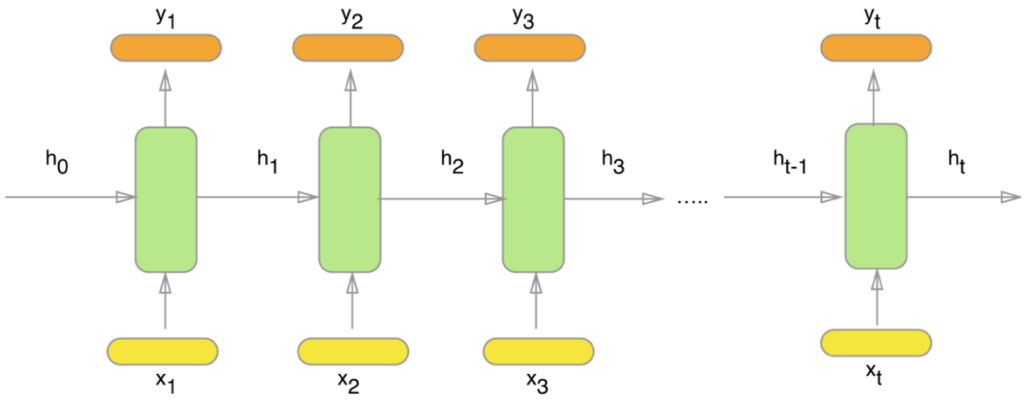
Software/Package	**Use**
Python 3.7	For code
TensorFlow	Provides backend for model
SKLearn	For Validation.
Pandas	For character conversions and calc.
NumPy	For numeric calc.
Matplotlib	For plotting graphs.

3.2. Solutions Approach

In this project , we have decided to use Gated Recurrent Unit (GRU) which is a variant of Recurrent Neural Network(RNN) architecture. First let's look at the working of RNN to better understand it's variants- LSTM and GRU.

RNN-

A Recurrent Neural Network is a mix of multiple feedforward neural networks which transfer information from one feedforward neural network to another and use internal state memory for processing sequences.



Recurrent neural network

As we see x1, x2, x3, .., xt depict the input words from the text, y1, y2, y3, .., yt depict the predicted words and h0, h1, h2, h3, .., ht hold the information for the previous words.

Below is a mathematical formulation of RNN:

$$h(t) = f_H\left(W_{IH}x(t) + W_{HH}h(t-1)\right)$$
$$y(t) = f_O\left(W_{HO}h(t)\right)$$

The W_{IH}, W_{HH}, and W_{HO} are the weight matrices and f_O and f_H are the output and hidden unit activation functions. Above we see x(t) and y(t) as input and output vectors.

Problems with a standard RNN-

The RNN model comes with a drawback, known as the **vanishing gradient problem,** which reduces its accuracy. To update a neural networks weights, gradient values are used. The vanishing gradient problem is due to the shrinking of the gradient value as the model back propagates. When gradient value becomes extremely small, it doesn't contribute much to learning and **the network experiences difficulty in learning words or info which are distant** and makes predictions only based on the most recent data. This is when different vaiants like the LSTM or the GRU become useful.

LSTM-

Long Short Term Memory(LSTM) is a variant of RNN with forward feedback connections.

A LSTM unit is comprises of a cell along with gates such as an input gate, an output gate and a forget gate, which allow the cell to memorise values for an random amount of time. The gates are responsible for controlling the flow of information through LSTM cell.

The LSTM's hidden state (h_t) can be calculated in the following manner:

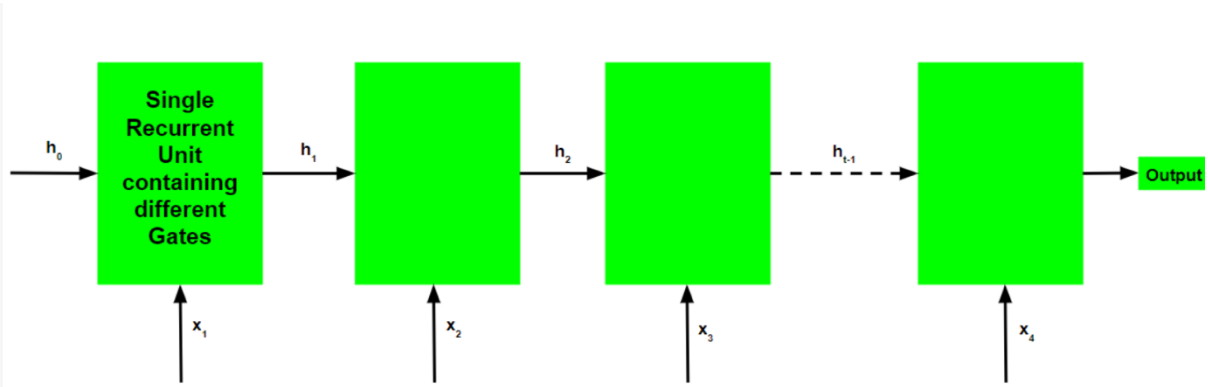
$$i_t = \sigma\left(x_t U^i + h_{t-1}W^i\right)$$
$$f_t = \sigma\left(x_t U^f + h_{t-1}W^f\right)$$
$$o_t = \sigma\left(x_t U^o + h_{t-1}W^o\right)$$
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1}W^g\right)$$
$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$
$$h_t = \tanh(C_t) * o_t$$

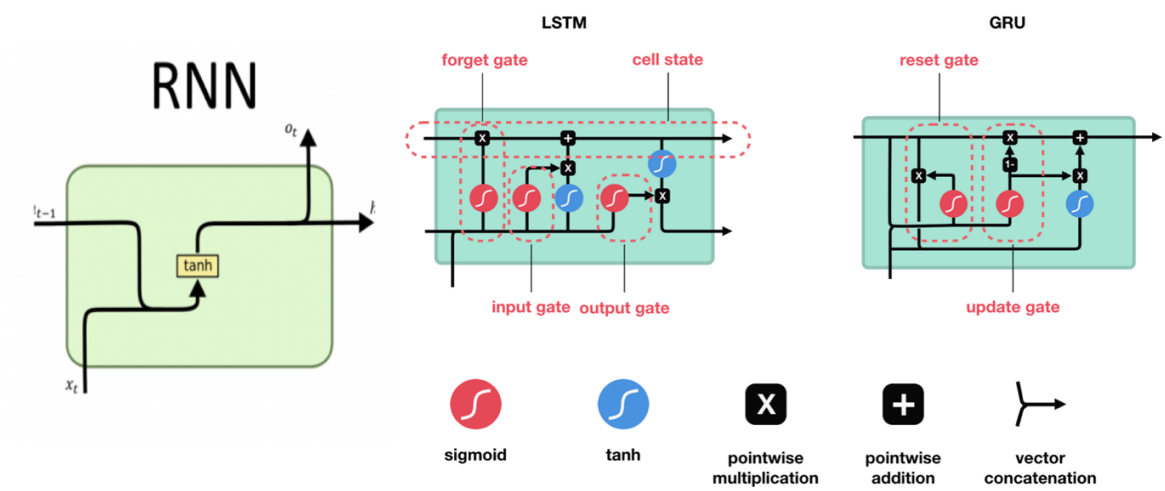
The variables, **i,o,f** are the input, output and forget gates, respectively. **W** is used to denote the recurrent connection from the preceding hidden layer to present hidden layer whereas **U** denotes weight matrix connecting the inputs to the present hidden layer.

GRU-

The GRU formally known as the Gated Recurrent Unit is a variant of RNN .The GRU has gates such as the reset gate and update gate and current memory gate to tackle the problem of vanishing gradient. The different gates are described as follows:

1. **Update Gate(z):** It tells how much of the previous information needs to be passed to the future.
2. **Reset Gate(r):** It tells how much of the previous information to remove or forget.





Simple RNN VS LSTM VS GRU

Job of the reset gate is to tell how to combine the new input with the previous memory data whereas the update gate defines how much of the previous memory data to keep in memory.

The GRU's hidden state h_t can be calculated in the following manner:

$$z_t = \sigma\left(x_t U^z + h_{t-1} W^z\right)$$
$$r_t = \sigma\left(x_t U^r + h_{t-1} W^r\right)$$
$$\tilde{h}_t = \tanh\left(x_t U^h + (r_t * h_{t-1}) W^h\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

The ***r*** stands for the reset gate whereas ***z*** stands for update gate. GRUs have performed better than LSTMs while also being faster due to a less complex architecture

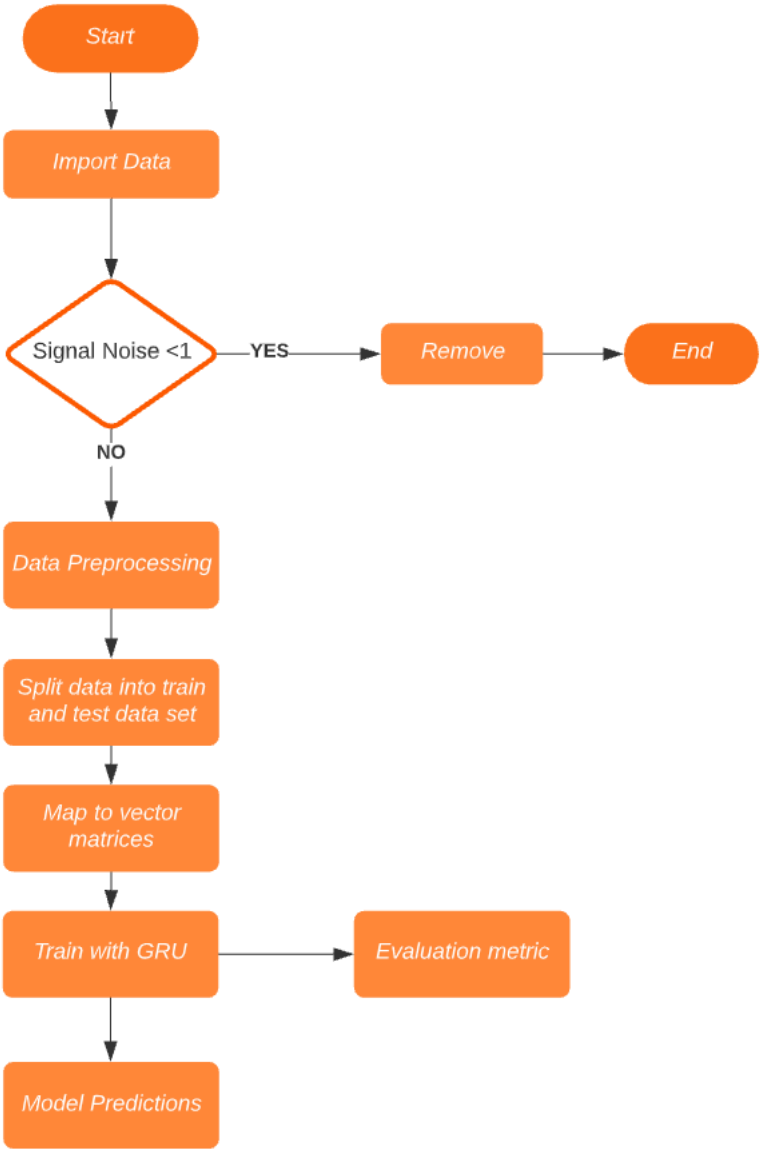
BIDIRECTIONALISM(BI-GRU)

RNN model information flows in single direction. Research has indicated that by reversing the sequence i.e the model sequences in both directions (both .. y_{i-2} , y_{i-1} , y_i . . . and...y_{i+2},y_{i+1},y_i) ,performance is improved. Bidirectional modelling can be done by implementation of a bidirectional structure inside the framework of RNN , or two RNN working in reverse directions can be fused to achieve the same goal. The latter has been used in our work.

4. MODELLING AND IMPLEMENTATION DETAILS

4.1. Design Diagrams

4.1.1. Flow diagram



High level diagram of the overall model architecture.

4.2. Implementation details and issues

Dataset-

This project made use of a total of 6034 RNA sequences for testing the models. The training package used 3029 of these sequences with a length of 107 nucleotides. A study dataset containing 3005 sequences, each of which had a length of 130 nucleotides, was employed. The sequences which are 107 character long have 68 scored positions whereas those having length of 130 have 91.

Using three predictors to determine the likelihood of a sequence, we can forecast each sequence based on: the nucleotide sequence itself (expressed in A, G, C, and U nucleotides), the molecule's projected structure, and the Predicted Loop Form. Additional features include each unique sequence as well as a Base Pair Probability Matrix, which details the probability of various base pair interactions.

Values of reactivity, degrad. at pH 10, degrad. values at pH 10 with added magnesium, and degradation values at 50 were also given for the first 68 base pairs of sequences with length 107 and first 91 base pairs for those with length 130

Data Pre-processing-

1. Loading Data And Analysis-

We are provided with JSON object for each sample. Each sample contains

1. an RNA sequence represented as A, G, U or C characters.

	sequence
0	G
1	G
2	A
3	A
4	A
...	...
102	A
103	C
104	A
105	A
106	C

```
'.....((((((.......)))).)).((.....((..((((((....))))))..))....))....
(((((((....)))))))....................'
```

2. metadata for each sequence element, including the structure property, which represents if it's "paired" or "unpaired", and predicted loop type, which provides an estimate of the "loop type".
3. In the public train/test set the sequence is 107 long as compared to 130 long in the private test set.
4. Structure of the sequence is described using "(" and ")" for paired and "." for unpaired.

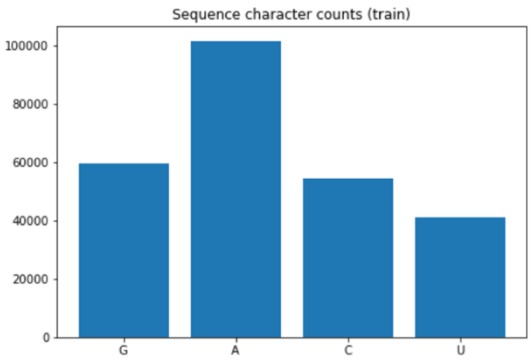
5. There are different loop types such as X, E, I, B, H, M AND S

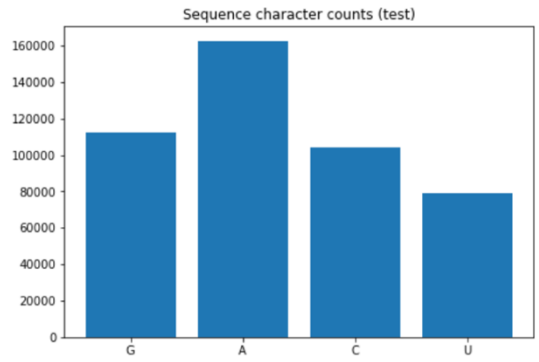
6. There are 5 labels that need to be predicted for each sample in test set :
    - Reactivity
    - deg_Mg_pH10
    - deg_pH10
    - deg_Mg_50C
    - Deg_50C

2. IN DEPTH EDA

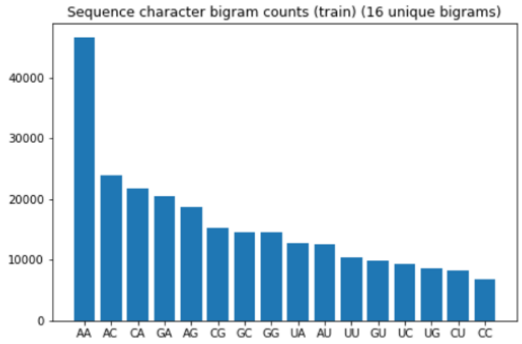
    A. Sequence-

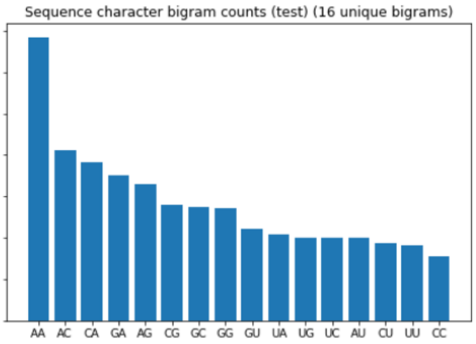
- We'll explore the frequency of each sequence character (A, G, U, C) across all samples (Train and Test).



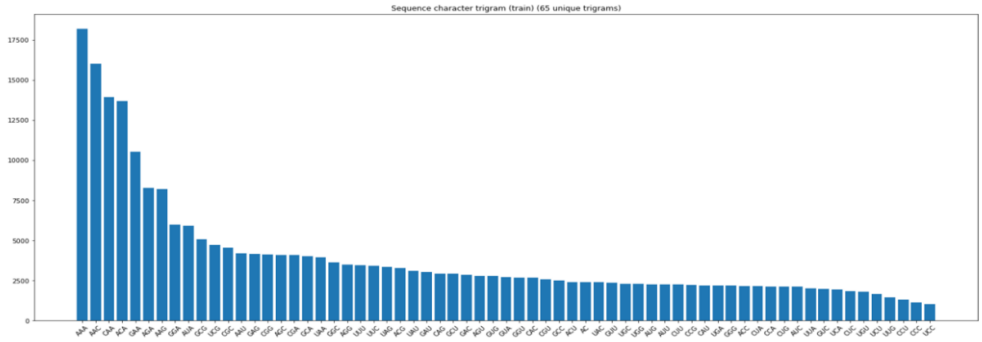


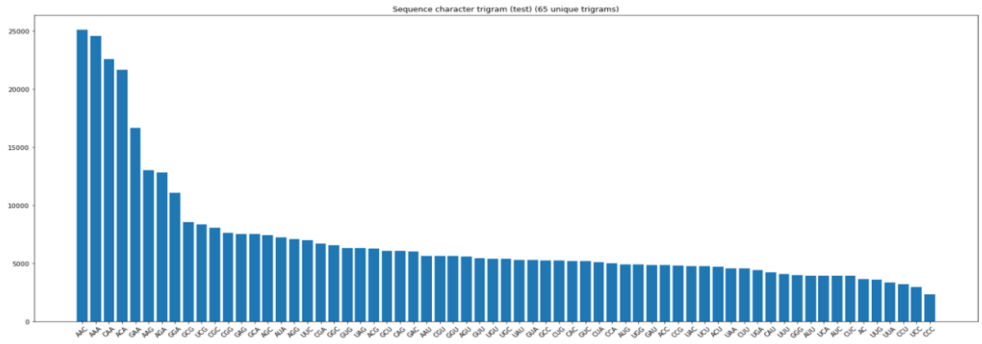
- Now we are interested in learning about how frequently certain characters appear side-by-side. To do that, we're going to break the sequences into bigrams and trigrams and explore the frequency.





- Above we can see the frequency of different Bigrams.

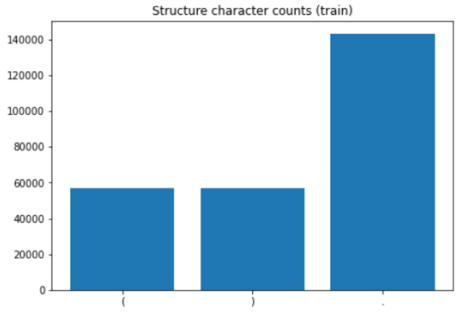


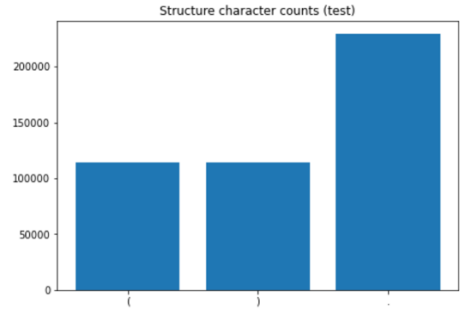


- Above we have frequency of trigrams.

B. Structure

- We'll explore the frequency of each structure.

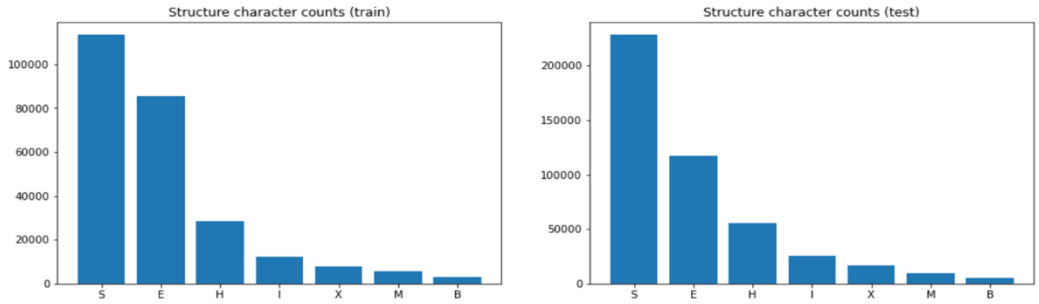




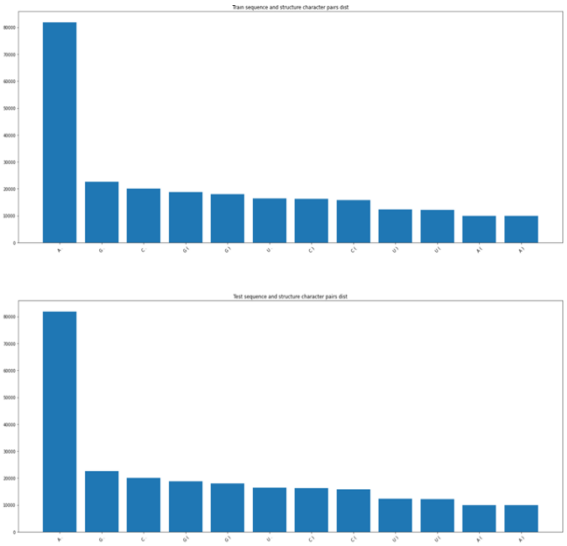
- We can notice that "(" and ")" are equal in number which gives rise to the conclusion that they are parseable "(" opens and ")" closes the pairs.

Predicted Loop Type

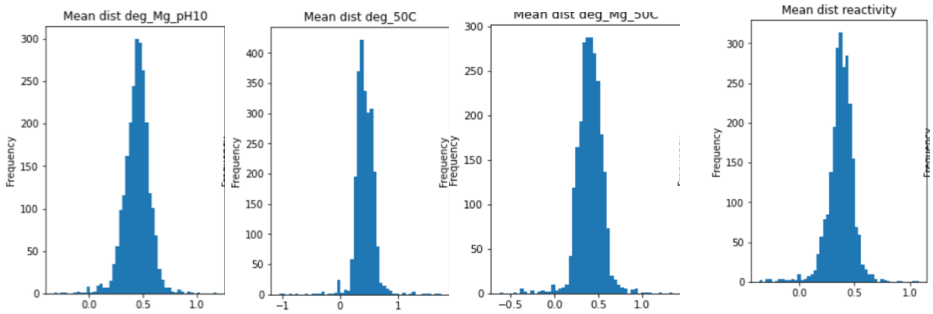
- Now we'll have a look at the frequency of different predicted loop types



- Now let's see how predicted loop relates to the sequence character



- Lets explore the different parameters- deg_Mg_pH10 , reactivity ,deg_50C, deg_Mg_50C 's mean distribution.



Data-preprocessing-

1. **Filtering Noise-** We will filter out records with noise less than 1.0 as required by dataset.

```
train = train.query("signal_to_noise >= 1")
train.shape
```

2. **Function To Convert Variable To Array-** The function pandas_list_to_array(df) would help us converting the target variables into an array which can be fed into the keras model.

```
def pandas_list_to_array(df):
    """
    Input: dataframe of shape (x, y), containing list of length l
    Return: np.array of shape (x, l, y)
    """

    return np.transpose(
        np.array(df.values.tolist()),
        (0, 2, 1)
    )
```

3. **Conversion Of Character To Integer-** The function preprocess_inputs(df, token2int, cols=['sequence', 'structure', 'predicted_loop_type']) Takes care of conversion using dictionary(token2int) to convert character to integer

```
def preprocess_inputs(df, token2int, cols=['sequence', 'structure', 'predicted_loop_
    return pandas_list_to_array(
        df[cols].applymap(lambda seq: [token2int[x] for x in seq])
    )
```

4. **Helper Function For Evaluation Metric-** The reason we are using MCRMSE in this challenges is because there are multiple outputs that we are trying to predict.
   Normally, we can calculate RMSE to get a single-number evaluation metric for our prediction, but if we are predicting multiple values at once—in the case of the OpenVaccine competition, we need to predict degradation rates under multiple conditions—we would get multiple different RMSE values, one for each column.

The MCRMSE is simply an average across all RMSE values for each of our columns, so we can still use a single-number evaluation metric, even in the case of multiple outputs.

Model Implementation-

1. We have used Bi-directional GRU for prediction of values of various predictor columns. Building of the model is shown below-

```
Model: "functional_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 107, 3)]          0
_________________________________________________________________
embedding (Embedding)        (None, 107, 3, 200)       2800
_________________________________________________________________
tf_op_layer_Reshape (TensorF [(None, 107, 600)]        0
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 107, 600)          0
_________________________________________________________________
bidirectional (Bidirectional (None, 107, 512)          1317888
_________________________________________________________________
bidirectional_1 (Bidirection (None, 107, 512)          1182720
_________________________________________________________________
bidirectional_2 (Bidirection (None, 107, 512)          1182720
_________________________________________________________________
tf_op_layer_strided_slice (T [(None, 68, 512)]         0
_________________________________________________________________
dense (Dense)                (None, 68, 5)             2565
=================================================================
Total params: 3,688,693
Trainable params: 3,688,693
Non-trainable params: 0
_________________________________________________________________
```

2. For bi-GRU networks ,individual networks were duplicated after which they were made to work in opposite direction and fused with the initial forward networks by concatenating th eir outputs. It has 3 layers and has a dropout.

```
def build_model(embed_size, seq_len=107, pred_len=68, dropout=0.5,
                sp_dropout=0.2, embed_dim=200, hidden_dim=256, n_layers=3):
    inputs = L.Input(shape=(seq_len, 3))
    embed = L.Embedding(input_dim=embed_size, output_dim=embed_dim)(inputs)

    reshaped = tf.reshape(
        embed, shape=(-1, embed.shape[1],  embed.shape[2] * embed.shape[3])
    )
    hidden = L.SpatialDropout1D(sp_dropout)(reshaped)

    for x in range(n_layers):
        hidden = gru_layer(hidden_dim, dropout)(hidden)


    truncated = hidden[:, :pred_len]
    out = L.Dense(5, activation='linear')(truncated)

    model = tf.keras.Model(inputs=inputs, outputs=out)
    model.compile(tf.optimizers.Adam(), loss=MCRMSE)

    return model
```

3. The implementation of Bi-directional GRU has been completed. The different hyperparam eters like the dropout values ,number of layers , their size and embedding size were tested and the lowest values were selected after their performances were stagnant. Epoch value o f 50 has been selected along with adam optimizer and sigmoid activation function and loss metric as MCRMSE.

```
Epoch 48/50
59/59 - 4s - loss: 0.1727 - val_loss: 0.2165
Epoch 49/50
59/59 - 4s - loss: 0.1721 - val_loss: 0.2159
Epoch 50/50
59/59 - 4s - loss: 0.1716 - val_loss: 0.2162
```

4. Values of predictor values are predicted

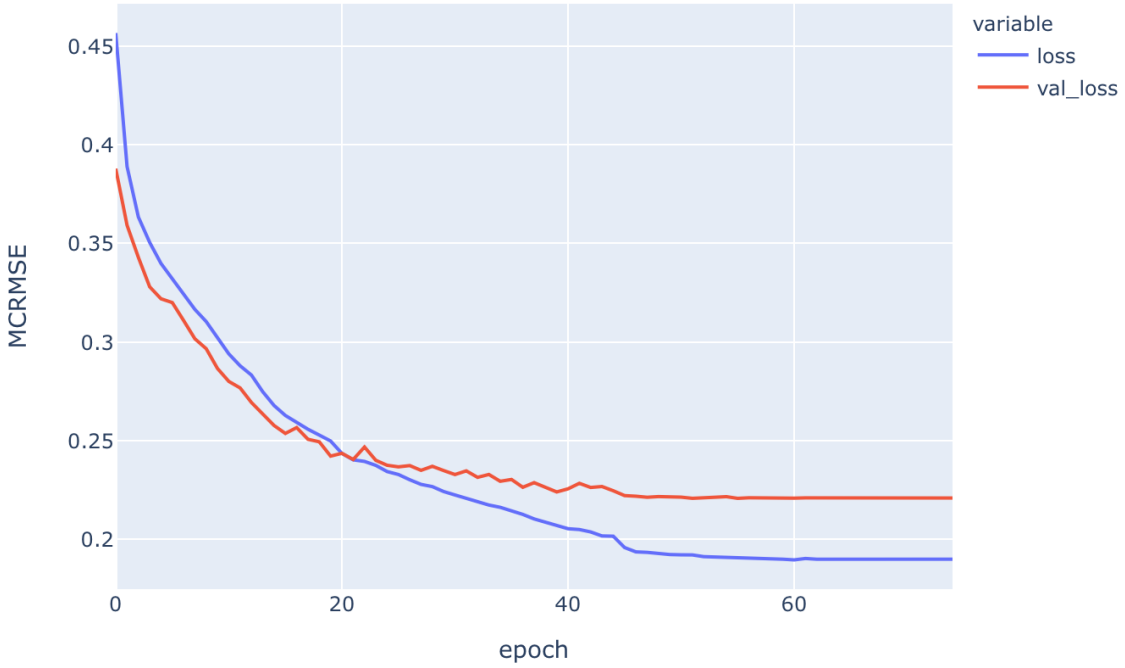
	reactivity	deg_Mg_pH10	deg_Mg_50C	deg_pH10	deg_50C	id_seqpos
0	0.708256	0.659190	0.546342	2.034063	0.757310	id_00073f8be_0
1	2.104239	3.170810	3.394956	4.254514	2.986442	id_00073f8be_1
2	1.610737	0.718315	0.724157	0.650183	0.793168	id_00073f8be_2
3	1.344301	1.275968	1.805824	1.216240	1.918668	id_00073f8be_3
4	0.858349	0.669474	0.924544	0.475049	0.849421	id_00073f8be_4

5.RESULTS, CONCLUSION, AND FUTURE WORK

5.1. Results & Conclusion

Through our project we've predicted values of rate of degradation at different locations along the RNA sequence and calculated the training loss=0.1716 and validation loss=0.2162.

The difference in the gap is the generalization gap. Through the predictions, newer sequences have been validated and can be used for building stable RNA sequences which will help in the production vaccines.



5.2. Future Work

The research for mRNA vaccines to fight future insurgence by pathogens is crucial. We hope by helping in reducing mRNA degradation time we can prepare for the future pandemics in an efficient way.

The future step here would be to implement such models as ours on a greater dataset so that we can validate greater number of robust RNA sequences which do not degrade to reduce the scope of error to the smallest possible unit. We hope to continue the research to help the community fight future pandemics.

6. REFERENECES

1. Kipf, T., Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv (2017).
2. Pardi, N., Hogan, M. mRNA vaccines — a new era in vaccinology. NatRev Drug Discovery 17, 261–279 (2018).
3. Wayment Steele, H., Soon Kim, D., et al. Theoretical basis for stabilizing messenger RNA through secondary structure design. bioRxiv (2020).
4. Zhang, C., Maruggi, G., et al. Advances in mRNA Vaccines for Infecious Diseases. Fronteirs in Immunology (2019).
5. Jieliang Wang, Ying Peng, Haiyue Xu,Zhengrong Cui,and Robert O. Williams . The COVID-19 Vaccine Race: Challenges and Opportunities in Vaccine Formulation (2020).
6. Li, X.; Zhang, W.; Ding, Q. Cross-Domain Fault Diagnosis of Rolling Element Bearings Using Deep Generative Neural Networks. IEEE Transactions on Industrial Electronics 2019
7. Schmidhuber, J. Deep learning in neural networks: an overview. Neural Networks 2015
8. Liu, H.; Zhou, J.; Zheng, Y.; Jiang, W.; Zhang, Y. Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. ISA Trans. 2018
9. Qiu, D.; Liu, Z.; Zhou, Y.; Shi, J. Modified Bi-Directional LSTM Neural Networks for Rolling Bearing Fault Diagnosis. IEEE International Conference on Communications (ICC) 2019,
10. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. IEEE Transactions on Industrial Electronics 2018
11. Cho, K., Bahdanau, D, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.()
12. Hochreiter S., Schmidhuber J. Long Short Term Memory. Neural Computation 9(8):1735-1780 (1997).
13. P. Kingma, D.; Lei Ba, J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2015
14. Rong, M.; Shi, H.; Tan, S. Large-Scale Supervised Process Monitoring Based on Distributed Modified Principal Component Regression. Ind. Eng. Chem. Res. 2019
15. Tidriri, K.; Chatti, N.; Verron, S.; Tiplica, T. Bridging data- driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. Annual Reviews in Control 2016
16. Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

17. Cho, K.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078, 2014
18. G. L. Z. V. D. M. L. &. W. K. Q. Huang, " Densely connected convolutional networks.," CVPR, vol. 1, p. 3, 2017.
19. Ickwon Choi,Amy W. Chung,Todd J. Suscovich,Supachai Rerks-Ngarm,Punnee Pitisuttithum,Sorachai Nitayaphan,Jaranit Kaewkungwal,Robert J. O'Connell,Donald Francis,Merlin L. Robb,Nelson L. Michael,Jerome H. Kim,Galit Alter,Margaret E. Ackerman,Chris Bailey-Kellogg. Machine Learning Methods Enable Predictive Modeling of Antibody Feature. Journal- PLOS (2015).
20. G. H. A. K. I. S. R. S. Nitish Srivastava, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.
21. Kaggle