Northeastern

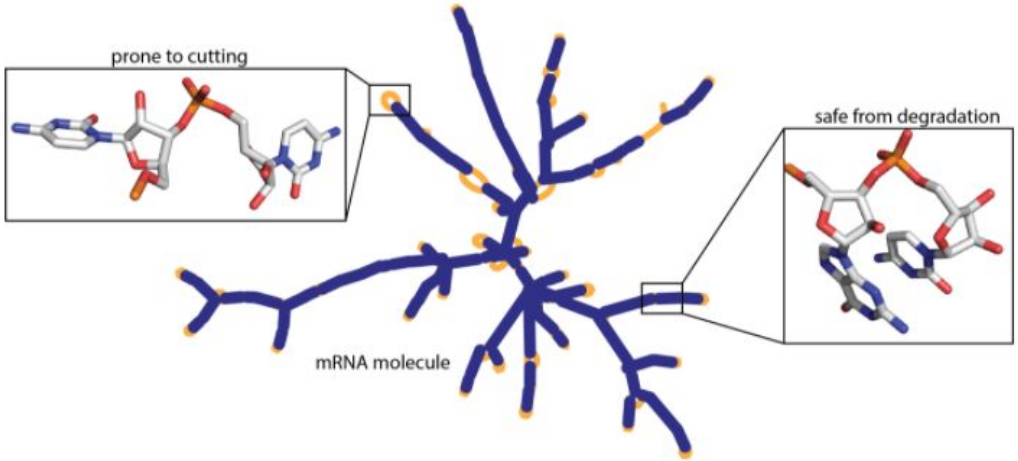
# Final Project Presentation

## COVID-19 mRNA Vaccine Degradation Prediction

CS7140 Advanced Machine Learning
Rajesh Sakhamuru, Zhiyu Chen

# Project Introduction

- With the attack of COVID-19 worldwide, scentific community is taking full effort to develope mRNA vacine.
- mRNA vaccine is unstable, depending on its sequence structure.
- Stanford-covid dataset provides the stability of some mRNA sequence.
- Our task is to predict stability of the unknown sequence.



# Dataset Structure-Feature

| sequence | structure | predicted_loop_type | signal_to_noise | SN_filter | seq length | seq_scored |
|---|---|---|---|---|---|---|
| string | string | string | float | int | int | int |
| 107 | 107 | 107 | >0.0 | 0 or 1 | 107 | 68 |

| index | id | sequence | structure | predicted_loop_type | signal_to_noise | SN_filter | seq_length | seq_scored |
|---|---|---|---|---|---|---|---|---|
| 0 | id_001f94081 | GGAAAAGCUCUAAUAACAGGAGACUAGGACUACGUAUUUCUAGGUA... | .....((((((.....)))).)).((.....((.. ((((((..... | EEEEESSSSSSHHHHHHHSSSSBSSXSSIIIIISSIISSSSSSHHH... | 6.894 | 1 | 107 | 68 |
| 1 | id_0049f53ba | GGAAAAAGCGCGCGCGGUUAGCGCGCGCUUUUGCGCGCGCUGUACC... | ..... ((((((((((((((((((((....)))))))))).)))... | EEEEESSSSSSSSSSSSSSSSSSSSSSSSSHHHHSSSSSSSSSSBSSS... | 0.193 | 0 | 107 | 68 |
| 2 | id_006f36f57 | GGAAAGUGCUCAGAUAAGCUAAGCUCGAAUAGCAAUCGAAUAGAAU... | .....((((.((.....((((.(((.....))))).. ((((.....)... | EEEEESSSSISSIIIIISSSSMSSSHHHHHSSSMMSSSSHHHHHHS... | 8.800 | 1 | 107 | 68 |

# Dataset Structure-prediction

| | | |
|---|---|---|
| reactivity | float list | length:68 |
| deg_Mg_pH10 | float list | length:68 |
| deg_pH10 | float list | length: 68 |
| deg_Mg_50C | float list | length: 68 |
| deg_50C | float list | length: 68 |

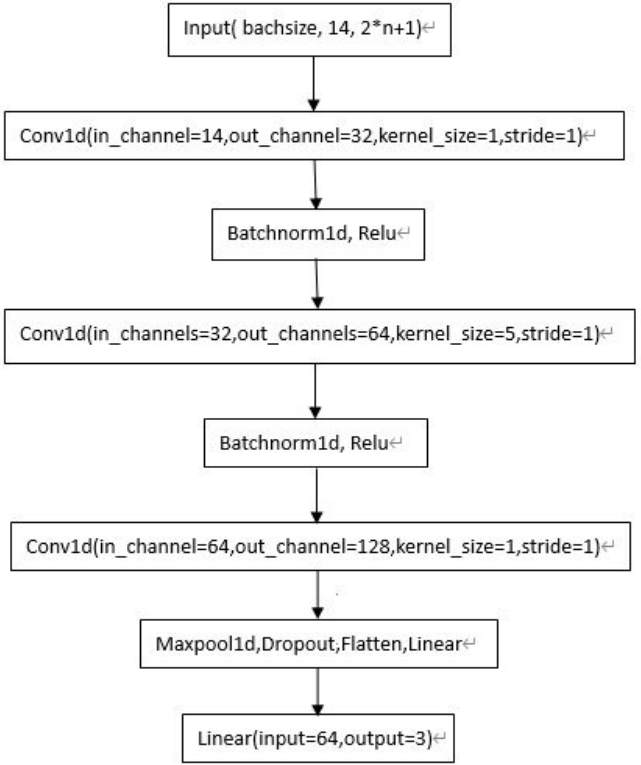
| | | |
|---|---|---|
| reactivity_error | float list | length:68 |
| deg_error_Mg_pH10 | float list | length:68 |
| deg_error_pH10 | float list | length:68 |
| deg_error_Mg_50C | float list | length:68 |
| deg_error_50C | float list | length:68 |

| reactivity_error | deg_error_Mg_pH10 | deg_error_pH10 | deg_error_Mg_50C | deg_error_50C | reactivity | deg_Mg_pH10 | deg_pH10 | deg_Mg_50C | deg_50C |
|---|---|---|---|---|---|---|---|---|---|
| [0.1359, 0.20700000000000002, 0.1633, 0.1452, ... | [0.26130000000000003, 0.38420000000000004, 0.1... | [0.2631, 0.28600000000000003, 0.0964, 0.1574, ... | [0.1501, 0.275, 0.0947, 0.18660000000000002, 0... | [0.2167, 0.34750000000000003, 0.188, 0.2124, 0... | [0.3297, 1.5693000000000001, 1.1227, 0.8686, 0... | [0.7556, 2.983, 0.2526, 1.3789, 0.637600000000... | [2.3375, 3.5060000000000002, 0.3008, 1.0108, 0... | [0.35810000000000003, 2.9683, 0.2589, 1.4552, ... | [0.6382, 3.4773, 0.9988, 1.3228, 0.78770000000... |
| [2.8272, 2.8272, 2.8272, 4.7343, 2.5676, 2.567... | [73705.3985, 73705.3985, 73705.3985, 73705.398... | [10.1986, 9.2418, 5.0933, 5.0933, 5.0933, 5.09... | [16.6174, 13.868, 8.1968, 8.1968, 8.1968, 8.19... | [15.4857, 7.9596, 13.3957, 5.8777, 5.8777, 5.8... | [0.0, 0.0, 0.0, 2.2965, 0.0, 0.0, 0.0, 0.0, 0.... | [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ... | [4.947, 4.4523, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ... | [4.8511, 4.0426, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,... | [7.6692, 0.0, 10.9561, 0.0, 0.0, 0.0, 0.0, 0.0... |
| [0.0931, 0.13290000000000002, 0.11280000000000... | [0.1365, 0.2237, 0.1812, 0.1333, 0.1148, 0.160... | [0.17020000000000002, 0.178, 0.111, 0.091, 0.0... | [0.1033, 0.1464, 0.1126, 0.09620000000000001, ... | [0.14980000000000002, 0.1761, 0.1517, 0.116700... | [0.44820000000000004, 1.4822, 1.1819, 0.743400... | [0.2504, 1.4021, 0.9804, 0.49670000000000003, ... | [2.243, 2.9361, 1.0553, 0.721, 0.639600000000... | [0.5163, 1.6823000000000001, 1.0426, 0.7902, 0... | [0.9501000000000001, 1.7974999999999999, 1.499... |

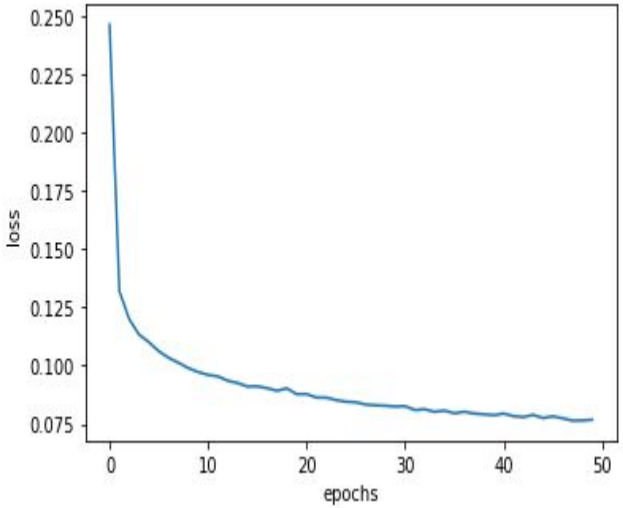
**Test sequence length: 107 or 130;  Prediction score length: 68 for public, 91 for private.**

# CNN model

- Assume stability of current base is only dependent on n previous and n afterwards bases
- Use one-hot encoding to encode current base
- Convolute 1d on sequence dimension
- Flatten and then use Dense layer to do prediction



# Result and Improvement



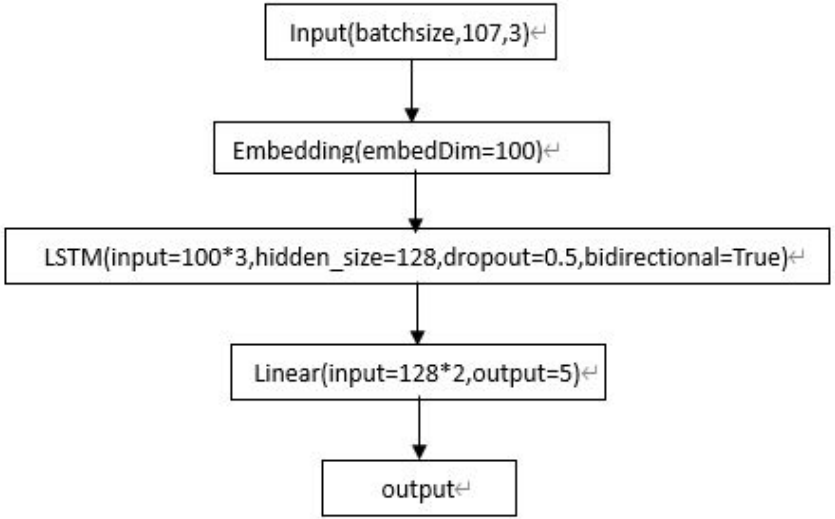

- Clean the data, only use high sign_to_noise data to train the model
- Learning embedding of for each sequence symbol, rather than use one-hot encoding
- Slight improvement likely due to the removal of noisy data for training.
- Embedding has the similiar effect of one-hot encoding

| Submission and Description | Private Score | Public Score |
|---|---|---|
| CNNSubmissionCONVPOOL (3).csv<br>a few seconds ago by Rajesh Sakhamuru<br>CNN Model Submission | 0.42497 | 0.32104 |

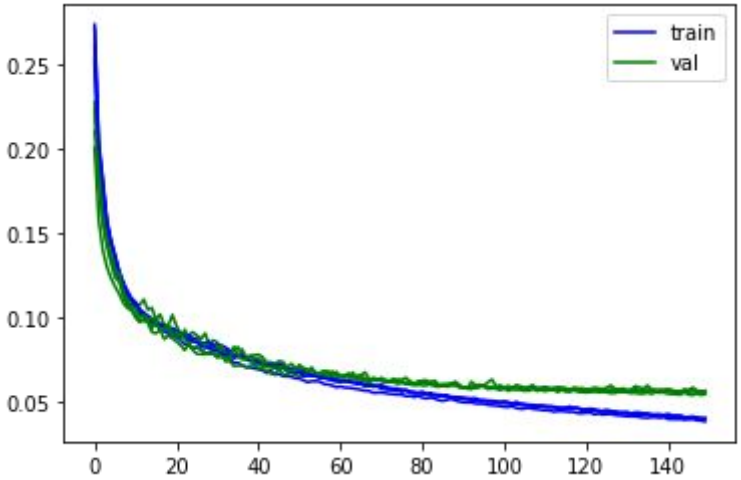
| Submission and Description | Private Score | Public Score |
|---|---|---|
| Emb_Clean_CNNSubmission.csv<br>a few seconds ago by Rajesh Sakhamuru<br>CNN with Embedding and de-noised data | 0.41448 | 0.32056 |

# LSTM Model

- Assume mRNA as time sequence sequence data
- The performance of  mRNA base is affected by both its previous bases and afterwards bases.
- Use the hidden layer of each base to predict the performance
- Split train and validation mutliple time to train multiple model to average prediction



# Result and improvement




- Bi-directional LSTM model performed significantly better than the CNN model.
- Has the ability to extract complex relationships at sequence locations to left and right of each nucleotide.
- One issue could be that the training data only has 68 predictions, the weights to predict after 68th prediction might not be updated in the training process.

Attempted Improvement:

- Add one more attention layer which could learn the affection weight of each time step to current time step.
- However, the result does not improve.

| Submission and Description | Private Score | Public Score |
|---|---|---|
| LSTMSubmission.csv<br>a few seconds ago by Rajesh Sakhamuru<br>LSTM Model | 0.38448 | 0.27012 |

| Submission and Description | Private Score | Public Score |
|---|---|---|
| ATT_LSTMSubmission.csv<br>a few seconds ago by jackychen718<br>add submission details | 0.38316 | 0.29657 |

# Reference

https://www.kaggle.com/c/stanford-covid-vaccine