Predicting New COVID-19 Mutations and Generating Predicted Sequences as Time Series Data

- The main goal of this project is to predict and generate new possible COVID-19 sequences. Suppose we have the RNA sequences of COVID-19 from the GISAID or NCBI gene banks of the last 3 years (Dec 2019 to Dec 2022, can work with only one year of data for now if it's too large). The first step of the model is to learn the possible mutation pattern from those 3 years, which can be treated as time-series data. The model has to predict the possible new mutations of the 4th year and generate those possible mutated sequences. This learning can be done using NLP and deep learning methods (LSTM, CNN, etc.).

- The next step would be to measure how strong the sequence is and what the probability is that a particular sequence will survive or not. And how valid the generated sequence is. This evaluation part can be done afterwards; for now, the first step is more important as it is an ongoing project.

1. **Any other probable approach regarding the goal is completely fine as long as it involves computational/programming approaches.**

2. **Some of the sampled experiment lists in a doc file provided by the task giver are attached separately for better understanding.**

There aren't many promising papers or existing works done till now. These approaches may come in handy for implementing the model:

1. https://www.hindawi.com/journals/mpe/2021/9980347/#B22 (Instead of influenza covid data will be used)
2. https://www.mdpi.com/2227-7390/10/22/4267