NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Antibody-Antigen Interaction

Prediction using transformer-based machine learning

Student: Cho Qi Xiang Supervisor: Kwoh Chee Keong

Project Objectives:

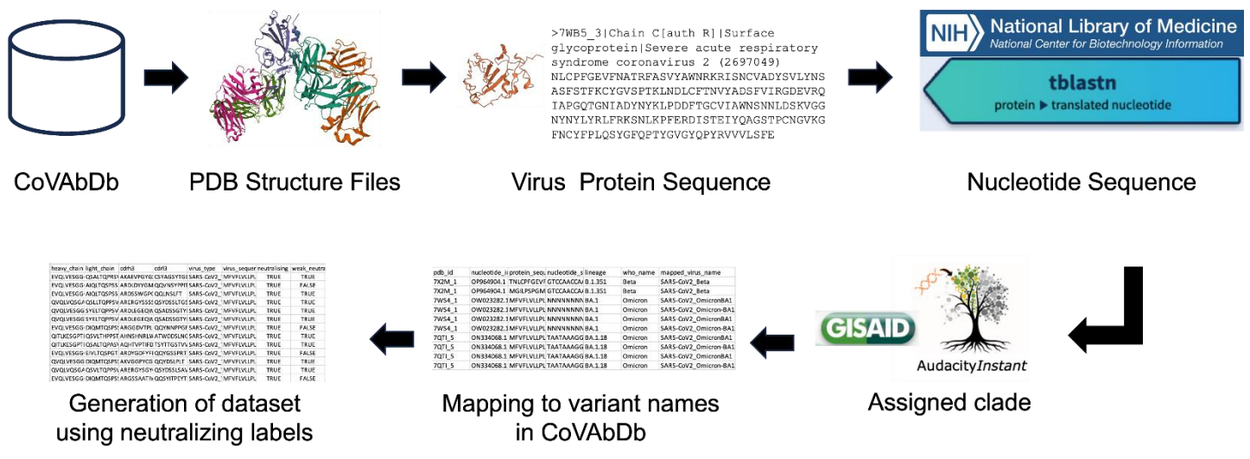
Use Natural Language Processing techniques and Machine Learning to predict neutralization classification of antibodies using textual representation of protein sequences

- Protein Sequences can be represented in textual format (FASTA Sequences)
- Each character in a FASTA Sequence corresponds to one of the 20 amino acids
- Protein sequences can be treated as sentences, and individual amino acids as words

Dataset & Data Preparation

Coronavirus Antibody Database

- 599 entries
- Sourced 835 FASTA sequence files
- Obtained nucleotide sequences using tBLASTn, then used to determine virus variants
- Generate dataset with 188k antibody-antigen sequence pairs



(A) Alphabetical Languages (e.g. English, French)

Text sequence: The lazy brown fox

Word-level tokenization: [*start*] [the] [lazy] [brown] [fox]

Character-level tokenization: [*start*] [t] [h] [e] [ ] [l] [a] [z] [y] [ ] [b] [r] [o] [w] [n]…

(B) Logographic Languages (e.g. Mandarin, Japanese Kanji)

Text sequence: 一个风和日丽的早上

Character-level tokenization: [*start*] [一] [个] [风] [和] [日] [丽] [的] [早] [上]

(C) Protein Sequences

Text sequence: NLCPFGEVFNA

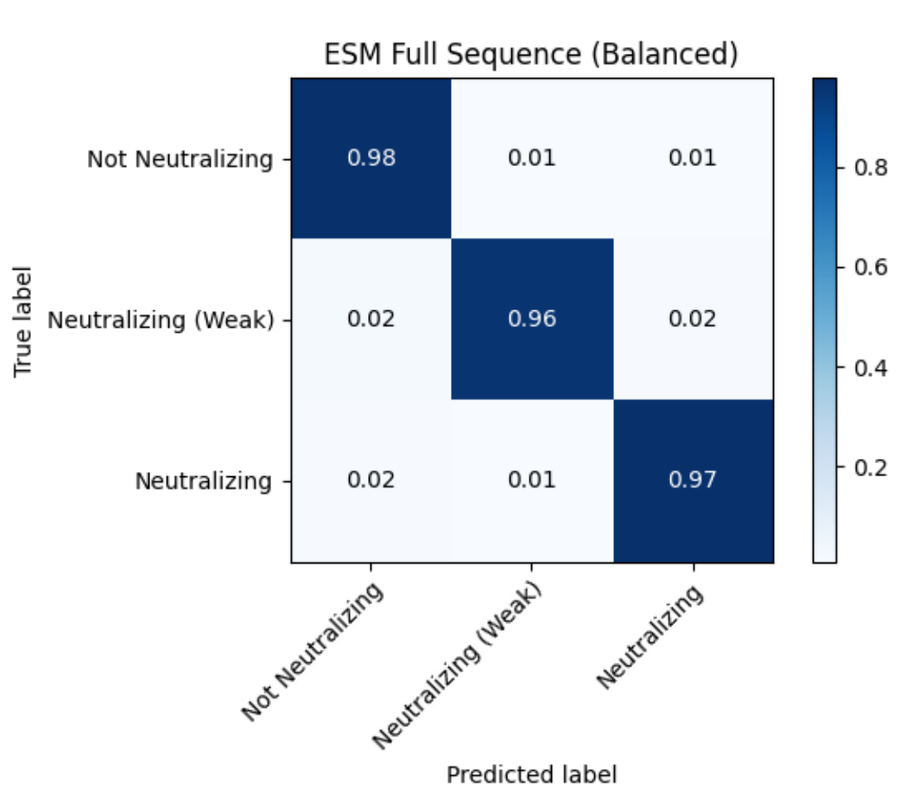
Character-level tokenization: [*start*] [N] [L] [C] [P] [F] [G] [E] [V] [F] [N] [A]

Models used

- **Logistic Regression** fitted with Graph Featurization with mean pooling
- **ESM2 Pre-trained protein language model** with character-level tokenization

Results: 98.3% Accuracy

- **94.72%** on a undersampled dataset for class imbalance
- **Using full antibody sequences** beneficial as compared to CDR3 region



Use cases & Future Work

- **Efficient validation** of neutralizing properties for engineered antibodies
- Extrapolate to to predict **neutralization ranges** using **IC50 values**