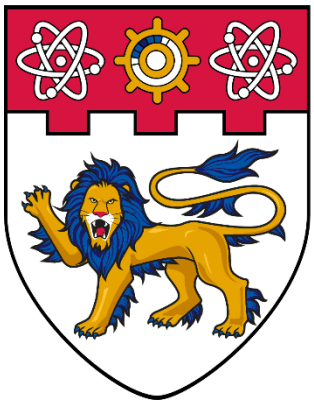

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

SCSE22-0968

PREDICTING ANTIBODY-ANTIGEN INTERACTIONS WITH TRANSFORMER-BASED MACHINE LEARNING

CHO QI XIANG (U2022896A)

SUPERVISOR: DR. KWOH CHEE KEONG

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Submitted in Partial Fulfillment of the Requirements for
the Degree of Bachelor of Engineering (Computer Engineering/Science)
Nanyang Technological University

2023

ABSTRACT

The neutralizing properties between antibodies to a given antigen are of extensive interest to the immunology community, with use in a range of applications such as vaccine development, immunotherapy, and drug design. Existing computational methods such as Molecular Dynamic (MD) simulations are computationally expensive, and require the crystal structures of these proteins which may not always be available.

Machine Learning (ML) is being introduced as an alternative to traditional computational methods. Supported by Deep Learning algorithms, work in the past has successfully demonstrated the feasibility of using Natural Language Processing (NLP) techniques in the antibody discovery process. The advent of transformer-based ML models is quickly replacing existing Deep Learning techniques, and has been used to predict paratope locations within antigens. The introduction of transfer-learning, where ML models trained on a task are adapted or fine-tuned to another related task has also given rise to Pre-trained Protein Language Models (PPLM). However, there is little work done on predicting the neutralizing properties between an antibody and an antigen using transformer-based ML models.

This project aims to demonstrate the viability of using transformer-based ML models to predict the neutralization properties between antibodies and antigens, focused on the SARS-CoV2 virus by fine-tuning a PPLM based on the Robustly Optimized BERT Approach (RoBERTa) architecture, to predict the neutralization properties between an antibody and an antigen, focused on the SARS-CoV-2 virus.

ACKNOWLEDGEMENTS

The author would like to express his appreciation and gratitude to his supervisor, Dr. Kwoh Chee Keong and research fellow Dr. Shamima Rashid for their guidance, support, and patience throughout the duration of the project.

Their knowledge, expertise, and willingness to provide constructive feedback and suggestions were invaluable to the completion of the project.

The author also gratefully acknowledges all data contributors, i.e. the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

An essential part of the human immune system is to recognize and neutralize foreign agents such as viruses, also known as antigens, to protect the body against infections and diseases. This is done through the production of antibodies, which are Y-shaped proteins that neutralize an antigen by binding to them.

The part of an antibody which binds to an antigen's receptor is a paratope, which is often found in the Complementarity-Determining Region 3 (CDR3), whereas the part of the antigen that binds to the antibody is an epitope. An example of an antibody-antigen complex with the antibody's heavy and light chain, as well as the virus spike protein sequence can be seen in Figure 1 below.

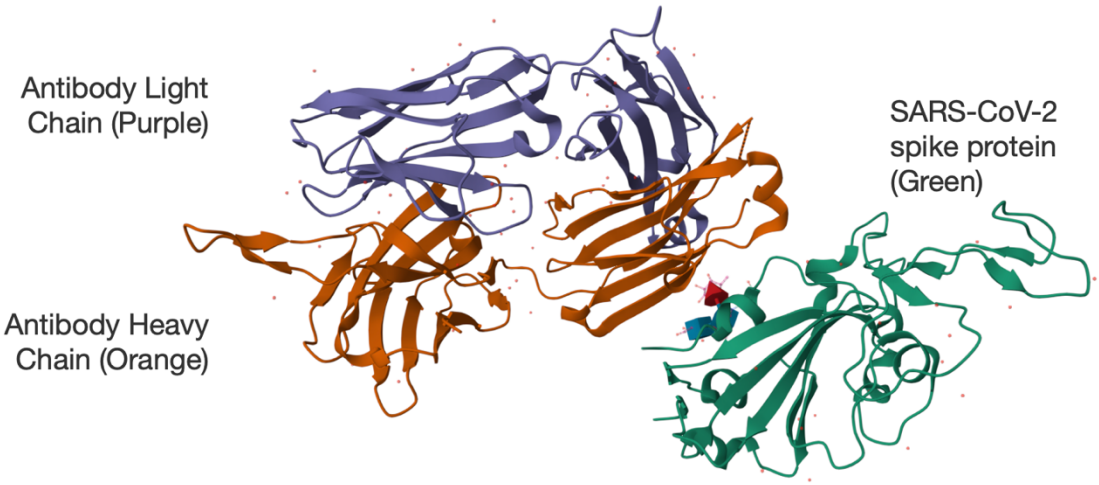

Figure 1: 3-Dimensional (3D) view of the 7T7B antibody-antigen complex, the heavy chain of the antibody is in purple, the light chain is in orange, and the SARS-CoV2 spike protein is in green [1].

The identification of antibody-antigen interactions has been used in vaccine development, and immunotherapy to treat cancers, and other critical illnesses. While computational methods are available for the screening of potential antibodies, these methods are computationally expensive and time consuming [2].

1.2 Machine Learning in Antibody Discovery

Machine Learning (ML) has been introduced to assist in antibody discovery. This is supported by a growing amount of biological data like protein sequences and structures, which are publicly available, and can be used to train such ML models. Magar et. al. demonstrated the possibility of using graph featurization and ML models like Logistic Regression to identify potentially neutralizing antibodies for SARS-CoV-2[3].

This method was replicated on a dataset consisting of 310 pairs of paratopes and epitopes. Out of these sequences, 228 were neutralizing, and 82 were not neutralizing. The results showed that Logistic Regression using mean pooling performed the best with an accuracy of 74.20%[4].

Deep Learning, a subfield of ML that uses algorithms modelled on the human brain's neural networks, has also been used in the prediction of antibody-antigen interactions. One such method is AbAgIntPre, which utilized a Siamese-like Convolutional Neural Network (CNN), trained on antibody-antigen sequence pairs. The results of this approach outperformed traditional ML models such as Random Forest and Logistic Regression used for a similar task[5].

1.3 Transformer models and Natural Language Processing techniques

Supported by Deep Learning algorithms, Natural Language Processing (NLP) techniques have been employed in antibody design. Protein sequences composed of amino acids can be represented in textual format, which are then embedded into a vector space[6]. Protein sequences can thus be treated as sentences, while the individual amino acids can be treated as words, as illustrated in Figure 2 below.

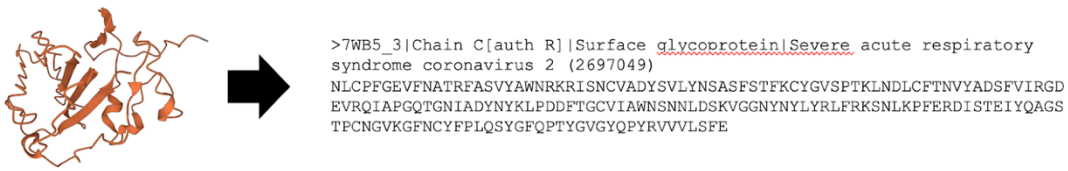

Figure 2: A SARS-CoV2 spike protein sequence, represented in the FASTA textual format, each character corresponds to one of the 20 amino acids (excluding rare or unconventional amino acids)[7].

Transformer-based models are quickly superseding other Deep Learning approaches and have become the dominant approach to NLP tasks. Transformer-based models have significant advantages over their other Deep Learning counterparts, such as better generalization with smaller amounts of data, owing to its self-attention mechanism[8]. AntiBERTa is a Bidirectional Encoder Representations from Transformers (BERT) Protein Language Model which has been used to predict the binding paratope locations from a B-cell Receptor, which is part of an antibody. The transformer-based approach performed better than other CNN-based models[9].

1.3 Transfer learning and Protein Language Models

Apart from the introduction of Transformer-based machine learning models, the introduction of transfer learning has been of immense benefit to the use of ML in solving various problems. Transfer learning is the process whereby a ML model that is trained on a task is adapted or fine-tuned for another related task, leveraging the knowledge and features learnt from its original task to improve the performance of the model for the targeted task.

Similar to how BERT language models have been created and fine-tuned for various tasks involving human languages[10], Pre-trained Protein Language Models (PPLM) and Pre-trained Antibody Language Models (PALM) have been introduced to understand protein sequences[11].

One such PPLM is the Evolutionary Scale Modelling (ESM) model. It is trained on a Masked Language Modelling (MLM) objective, which predicts randomly selected amino acids in a protein sequence, enabling it to learn their context with regards to the entire protein sequence. ESM is trained on approximately 250M individual protein sequences from the UniRef protein sequence database on a Masked Language Modelling (MLM) task[12].

The ESM model's architecture is also based on the Robustly Optimized BERT Approach (RoBERTa) architecture, similar to AntiBERTa. RoBERTa builds upon the fundamental ideas of BERT, and optimizes it to improve the pre-training process of the model[13].

While it has been shown that traditional ML and Deep Learning techniques have successfully been applied to predict antibody-antigen interactions, there is little research that explores the use of NLP and Transformer-based models to predict antibody-antigen interactions. Previous work has primarily shown such approaches used in the prediction of paratope locations.

1.4 Objective and Scope

This project presents an approach to predict antibody-antigen interactions using transformer-based ML models by fine-tuning a PPLM that is based on the RoBERTa architecture. The model was fine-tuned to a sequence classification problem, and was trained on a dataset consisting of 101,605 examples of protein sequences.

Each example consists of the protein sequences of the heavy and light chains of the antibody, the CDR3 sequences of the antibody, as well as the sequence of a spike protein of the SARS-CoV-2 virus, as well as their neutralizing information.

While the paratopes of the antibodies are frequently located within the CDR3 regions, proteins in nature often fold into complex three-dimensional structures, which is important to their biological functionalities. This means that representing the protein sequences in a textual format results in the loss of such information.

Hence, we also compare the usage of the full antibody sequences and the CDR3 sequences in the training of the model to determine whether the usage of the full sequences could provide additional context to the folding nature of these antibodies, allowing for a more accurate model.

The same datasets were also encoded and fitted on a Logistic Regression model, with results that show that the transformer-based approach significantly outperformed that of the Logistic Regression model.

CHAPTER 2: METHODOLOGY

2.1 Datasets

In order to train and test our models, the dataset from the Coronavirus Antibody Database (CoVAbDb)[14] was utilized. The dataset comprised 12,004 entries, with each entry containing (a) information about the published name of the antibody, (b) the antigens that the antibody have been proven to interact or neutralize with, (c) the sequence information which includes the antibody's heavy and light chains, as well as its highlighted CDR3 regions, (d) links to the primary literature of the antibody, (e) links to available structures which the antibody is involved in, and (f) the developmental origin of the antibody.

2.2 Data pre-processing and labelling

The sequence-based model aims to address a classification task to determine whether an antibody neutralizes a given antigen. To construct our input, we will need to curate a dataset with each entry comprising the sequence information of the heavy and light chains of the antibody, the spike protein sequence of the virus, as well as the neutralization classification.

As the CoVAbDb dataset did not include the original virus spike protein sequence information, we had to obtain these sequences from the referenced protein structures. We filtered for entries that contained structural references, which gave us a total of 599 entries.

The text in these entries containing the links to the protein structures hosted on the Protein Data Bank[15] was processed to obtain the protein structure IDs using a Regular Expression. Using the structure IDs, the protein sequences of these structures were downloaded in the FASTA format, which gave us 834 FASTA files consisting of the protein sequence information of the antibody's heavy and light chain, the virus spike protein, as well as any other relevant sequences. Each sequence also contains a description of the protein sequence in the header.

To extract the sequence information, along with its corresponding description, the BioPython package was used to parse and process the downloaded FASTA files[16].

The protein sequences within each FASTA file were iterated through, and classified into (a) antibody heavy chain, (b) antibody light chain, or (c) virus spike protein based on the corresponding description of the sequences.

Keywords in the sequence descriptions such as “light chain” and “heavy chain” were used to determine sequences within the structure that belonged to antibodies, whereas keywords such as “spike protein” were used to determine sequences that belonged to viruses. This allowed us to generate 990 entries consisting of the sequence information of the antibody’s heavy and light chain, as well as the virus’ spike protein sequences.

The neutralization information stored within the entries of CoVAbDb only mentions the name of the variant(s) of the virus which the antibody has been proven to neutralize. It was thus important to determine the variant of the virus spike proteins in order to accurately encode the neutralization properties between the antibody and the virus sequences. To identify the specific variants of the virus sequences, the translated nucleotide sequences of the spike proteins were obtained, and put through a web application to determine the specific lineage and variant[17].

To obtain the translated nucleotide sequence for the spike proteins, a search was done using tBLASTn on the Betacoronavirus database hosted on the National Centre for Biotechnology Information (NCBI)[18, 19]. Duplicate spike protein sequences were removed, and only sequences relevant to the SARS-CoV-2 were filtered out using the keyword “Coronavirus 2” on their descriptions, which resulted in a total of 273 spike protein sequences.

The tBLASTn result for each spike protein sequence with the lowest expectation value (e-value) was taken as the preferred nucleotide sequence translation. The e-value represents the expected number of random hits obtained by chance when querying the database for a specific sequence. As some of the virus spike protein sequences were too short, a total of 260 translated nucleotide sequences remained after performing tBLASTn.

GISAID’s AudacityInstant web application was used to determine the lineage and assigned World Health Organisation (WHO) variant names of the spike protein sequences[20]. The application takes in nucleotide sequences with a minimum of 250

bases and outputs the closest lineage, as well as the assigned WHO variant name, if applicable.

Duplicate values for the nucleotide sequences were dropped, and sequences with under 250 bases were filtered out, giving us a total of 78 nucleotide sequences. After the sequences were put through AudacityInstant, only 42 nucleotide sequences had a named WHO variant.

Using the lineage information, as well as assigned WHO variant names, the nucleotide sequences were manually mapped to the labels used in CoVAbDb. There were cases where the labels in the CoVAbDb dataset were more specific, such as some of the labels for the Omicron variant which included the specific lineage.

The nucleotide sequences were mapped to all relevant encodings. For example, two sequences of lineages BA.1.18 and BA.1.17.2 which were identified to be of the "Omicron" variant would be mapped to "SARS-CoV2_Omicron-BA1" due to the similarity in their lineages. These were then mapped back to their original pre-translated spike protein sequences, along with variant and lineage information. In addition, the reference spike protein sequence hCoV-19/Wuhan/WIV04/2019 was obtained from PDB, and mapped to the "SARS-CoV2_WT" label.

The entries in the CoVAbDb dataset were exploded to create an entry for each virus spike protein sequence the antibody neutralizes or does not neutralize against, and these were mapped to the various spike protein sequences for each variant based on the encodings used in the dataset. This resulted in a total of 188,038 entries, of which 98,127 entries are not neutralizing and 89,911 entries are neutralizing. Out of the 89,911 neutralizing entries, 24,875 entries had a weak neutralization property. An overview of the data pre-processing and labeling process is illustrated in Figure 3 below.

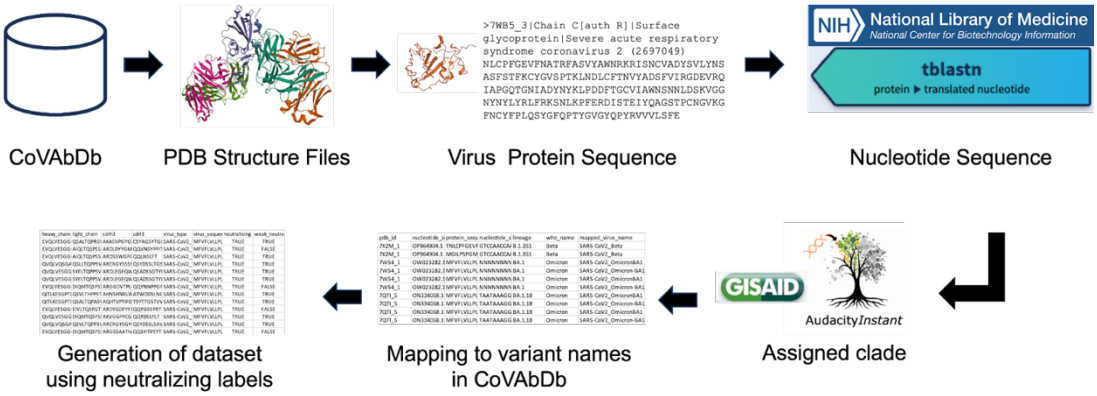

Figure 3: Overview of data cleaning, pre-processing and labelling process

2.3 Filtering usable data and tokenization

A key component of computational text analysis is tokenization, where text is split into individual tokens. In pre-trained Language Models designed for human languages, the approach chosen for tokenization varies across different languages.

For instance, in alphabetical languages such as English, individual words are typically used as tokens. Meanwhile, character-based languages, also known as logographic languages such as Mandarin, where there is no clear separation between words, character-level tokenization is often employed. Character-level tokenization may also be used for alphabetical languages, enabling flexibility for words that are unknown to the vocabulary or spelling errors.

As proteins do not have any well-defined vocabulary, tokenization at the word-level is not a well-defined option. Hence, the most common method for tokenization is character-level tokenization by treating individual amino acids as tokens[21], illustrated in Figure 4 below.

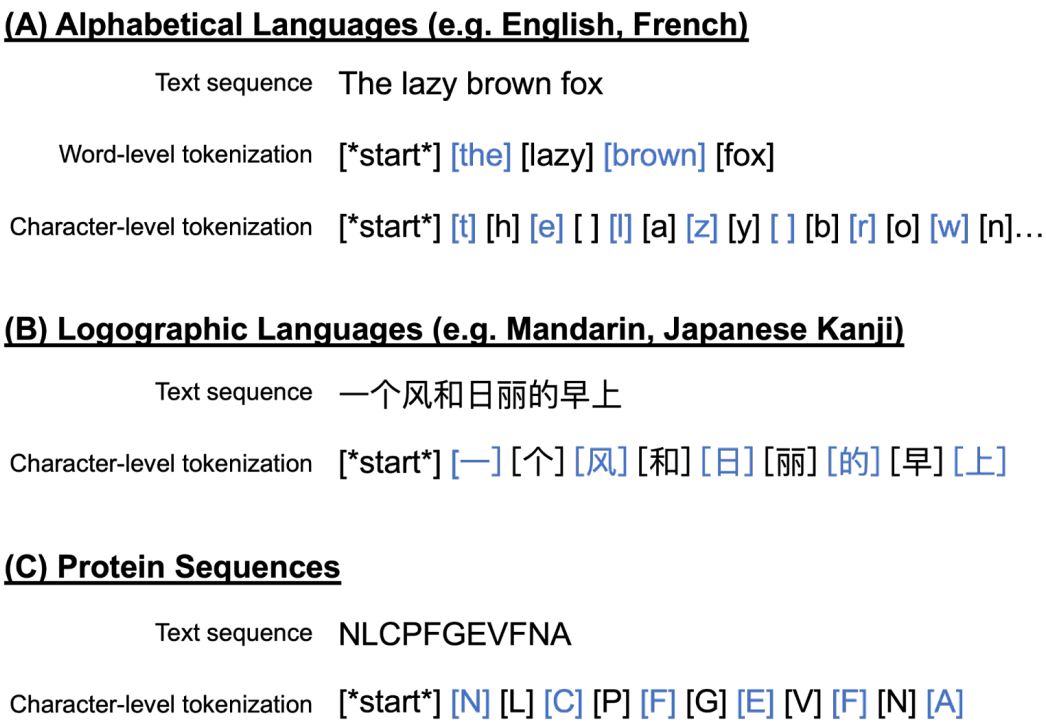

Figure 4: Some tokenization methods used for (a) alphabetical languages, (b) logographic languages, and (c) protein sequences.

Due to limitations in many PPLM model tokenizers, which use character-level tokenization, accepting sequences of a maximum of 1024 bases, or characters, examples where the total combined sequence length of the antibody heavy & light chain, and the virus sequence did not exceed 1024 characters were selected, allowing us to have a dataset of 101,605 entries. Of these entries, 56,568 were neutralizing, 13,822 had a weak neutralization, and 31,215 were neutralizing.

2.4 Resampling of data for class imbalance

After the pre-processing and encoding stage, the data was encoded into three classes. The examples in the dataset were labelled, where the entries which were not neutralizing were labelled “0”, weak neutralization labelled as “1”, and regular neutralization were labelled as “2”.

Noting that there was a class imbalance in the dataset, we performed undersampling on the overall dataset. This yielded us a balanced dataset consisting of a total of 41,466 examples with each class having 13,822 samples. This was done to prevent any potential biases within the model as a result of having a significantly larger number of

examples that are of the non-neutralizing class, and also due to the current size of the dataset.

2.5 Graph feature encoding of dataset for comparison and fitting of Logistic Regression Model

To compare the results of the transformer model approach, the model was compared against a similar problem trained on traditional ML techniques. We specifically compare against the methods used by Magar et al. where the protein sequences were encoded using graph featurization with mean pooling, and fitted on a Logistic Regression model[3].

We concatenate the antibody's heavy and light chains, as well as their CDR light and heavy chain sequences, and utilize the full virus sequence, then encode the examples using the same graph featurization technique.

Examples that were unable to be encoded using the aforementioned technique were omitted. A 75/25 train test split was then performed on the remaining examples. This process is repeated on both the imbalanced and balanced datasets.

The training data was fitted on the Logistic Regression model and the model's predictions were saved. Due to the significantly larger number of samples in the dataset, the maximum number of iterations on the Logistic Regression model was increased to 1,000.

2.6 ESM-2 PPLM Model training and finetuning

To perform our neutralization prediction task, we use the ESM-2 protein language model as its performance generally outperforms that of the original ESM model[22]. In addition, we utilize the HuggingFace's transformers library, which allows us to easily access and download ESM-2's model checkpoints[23]. The HuggingFace library is also matured, and is well known for its transformer library.

The "esm2_t12_35M_UR50D" checkpoint for ESM-2 was utilized. This checkpoint contains 12 layers and 35M hyperparameters. This smaller checkpoint was utilized to fit within our computational limitations.

The prediction problem was structured as a sequence classification problem, similar to classifying whether answers to a given question are correct, the input to the model included the antibody's protein sequences, the virus sequence, as well as the label corresponding to their neutralization classification.

To ensure fairness in comparison between the two methods, the model was trained using the same training data that was used to fit the Logistic Regression model, and validated using the same test data for both the balanced and imbalanced dataset.

As the ESM-2 model's tokenizer can only take in two protein sequences as input, the sequences for the antibody's heavy and light chains were concatenated. This was also done for the CDR3 sequences, similar to the approach used in the Logistic Regression method.

The sequences within the train and test sets were tokenized using the model's tokenizer, and the model was trained in a batch size of 8, for a total of 10 epochs. At the end of each epoch, the model was evaluated for its accuracy in correctly predicting the neutralization classification given an antibody and virus sequence.

At the end of the training, the best model among the epochs were selected for evaluation. This was always the model after the final 10th epoch. The predictions for the model were also saved.

2.7 Models trained and experiments conducted

In total we conducted four experiments comparing the accuracy of the fine-tuned ESM-2 model against the logistic regression model on both the imbalanced dataset prior to undersampling, as well as the balanced dataset, using the full antibody protein sequences, and the CDR3 sequences of the antibodies, shown in Table 1 below.

Table 1: Dataset and sequence used for training of the models for the four experiments, repeated on both the ESM-2 model and the Logistic Regression model

Dataset used	**Sequence used**
Imbalanced (101,605 examples) • 31,215 “Not Neutralizing” • 56,568 “Neutralizing” • 12,822 “Neutralizing (Weak)”	Full antibody sequence
	CDR3 Sequence
Balanced (41,466 examples) • 13,822 “Not Neutralizing” • 13,822 “Neutralizing” • 13,822 “Neutralizing (Weak)”	Full antibody sequence
	CDR3 Sequence

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Transformer-based approach outperforms Logistic Regression

3.1.1 Transformer approach is better for sequential data

We found that in general, the accuracy of the transformer based approach significantly outperformed the logistic regression model. The model was able to achieve an accuracy as high as 98.30% and 94.72% using the full antibody protein sequence on the imbalanced and balanced datasets respectively, compared to 60.20% and 44.50% using the Logistic Regression. The results are shown in Table 2 below.

Table 2: Accuracy of ESM2 transformer model and Logistic Regression model

Dataset & Model used	**Imbalanced multiclass dataset**		**Balanced multiclass dataset**	
	Full antibody sequence	*CDR3 only sequence*	*Full antibody sequence*	*CDR3 only sequence*
Logistic Regression	60.20%	59.77%	44.50%	44.70%
Transformer (ESM2)	98.30%	96.99%	94.72%	92.93%

The difference in model performance shows that the featurization used in the transformer approach could be better suited for datasets involving protein sequences, where the 3D crystal structures for the proteins are not available. This is likely due to the self-attention mechanism in the transformer approach, which allows it to capture the relationships between individual elements within the sequence.

In the context of predicting the neutralization properties of antibodies given the protein sequences, the model is able to learn the context for the individual amino acids represented within the FASTA sequences, and is hence able to perform better when the full antibody sequence is used potentially due to the additional context through the provision of the full sequence.

3.1.2 Using full antibody protein sequences improves model accuracy

While proteins are frequently considered linear polymers of amino acids, and can be represented in a string textual format composed of individual amino acids, the protein structures in reality fold into complex three dimensional structures which are not linear in nature. As mentioned earlier, these three-dimensional structures are critical in their biological functionality.

This suggests that the transformer approach is able to potentially learn the folding patterns and hence, potentially the biological functionalities of these sequences, which could have aided in the accuracy performance of the transformer model. The prediction of protein folding using the sequence representation alone through a transformer-based approach has also been shown in models such as ESMFold[21].

3.2 Post-prediction analysis

3.2.1 Transformer model accurately predicts minority classes on imbalanced dataset

To further validate the effectiveness of the transformer based approach, the predictions from the models were analyzed to check whether the minority classes were being accurately predicted despite being trained on an imbalanced dataset.

For the model which was trained on the full antibody sequences, it was found that the Logistic Regression method's predictions were heavily skewed towards the majority class, which was "Not Neutralizing". The true-positive rate for the minority class, which was "Neutralizing (Weak)" for the Logistic Regression was 0%, and 37% for the "Neutralizing" class.

Meanwhile, for the Transformer-based approach, it was found that it could achieve a high level of accuracy across all 3 classes despite being trained on an imbalanced dataset, achieving a true-positive rate of 99%, 95%, and 98% for the "Not Neutralizing", "Neutralizing (Weak)" and the "Neutralizing" classes respectively. However, this may also suggest a slight bias in the model towards the majority class as well. The confusion matrices for the ESM and Logistic Regression model are displayed in Figure 5.

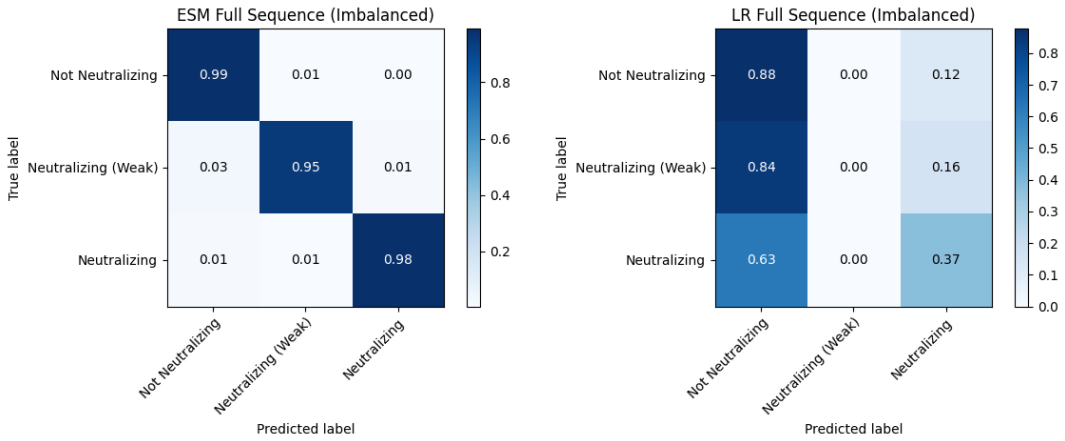

Figure 5: Confusion matrices for the models trained on the imbalanced dataset, using the full antibody sequences using (a) the ESM model, and (b) Logistic Regression

When trained using the CDR sequences, the ESM model approach achieves a true-positive rate of 98%, 92%, and 97% for the "Not Neutralizing", "Neutralizing (Weak)" and the "Neutralizing" classes respectively. This is slightly lower than when the model was trained on the full antibody sequences.

Meanwhile, the Logistic Regression approach continues to be heavily skewed towards the majority "Not Neutralizing" class, similar to the results when the full antibody sequence was used. The confusion matrices for the ESM and Logistic Regression model are shown in Figure 6.

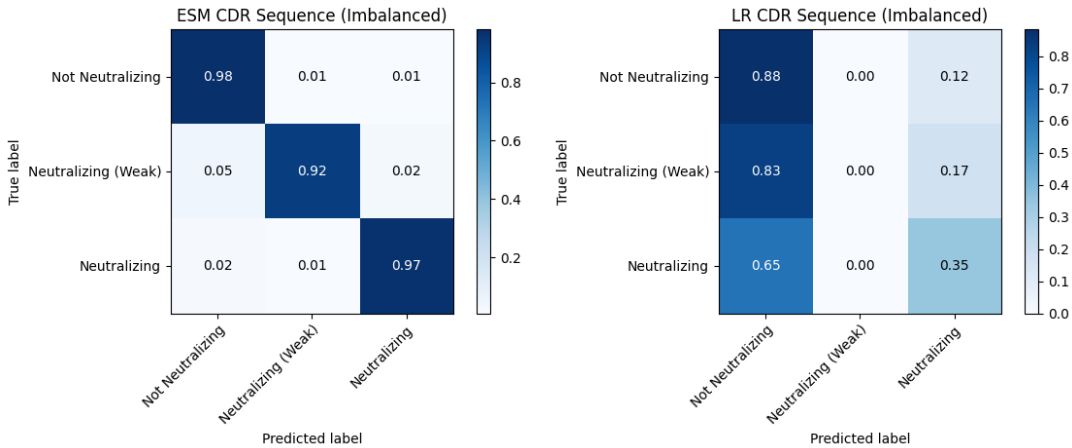

Figure 6: Confusion matrices for the models trained on the imbalanced dataset, using the CDR3 sequences using (a) the ESM model, and (b) Logistic Regression

3.2.2 Model performance is similar for balanced dataset despite lower accuracy score

In order to correct potential model bias from training on a biased dataset, the imbalanced dataset was undersampled as mentioned previously.

Despite achieving a slightly lower accuracy overall (94.72% for the balanced dataset as compared to 98.30% for the imbalanced dataset), we find that the minority "Neutralizing (Weak)" class is more accurately represented, achieving a higher true-positive rate of 96% as compared to the 95% previously.

Meanwhile, the other classes have a slight decrease in their true-positive rates, with 98% as compared to 99% previously for the "Not Neutralizing" class, and 97% as compared to 98% for the "Neutralizing" class.

For the Logistic Regression model, training it on a balanced dataset suggests a more accurate representation across the classes, despite the overall accuracy rate of 44.5% as compared to the 60.20% with the imbalanced dataset.

Training the Logistic Regression model on the balanced dataset shows that the previous minority class, which was the "Neutralizing (Weak)" class achieved an improved true-positive rate of 13%, as compared to 0% previously.

Despite this improvement, we note that the Logistic Regression method is still heavily skewed towards the "Not Neutralizing" class, even though the dataset had been fixed for class imbalance. This suggests that the Logistic Regression approach may not be able to support learning the complex relations between the amino acids within the protein sequences of the dataset. The confusion matrices for the models trained on the balanced dataset, using the full antibody sequences, is shown in Figure 7 below.

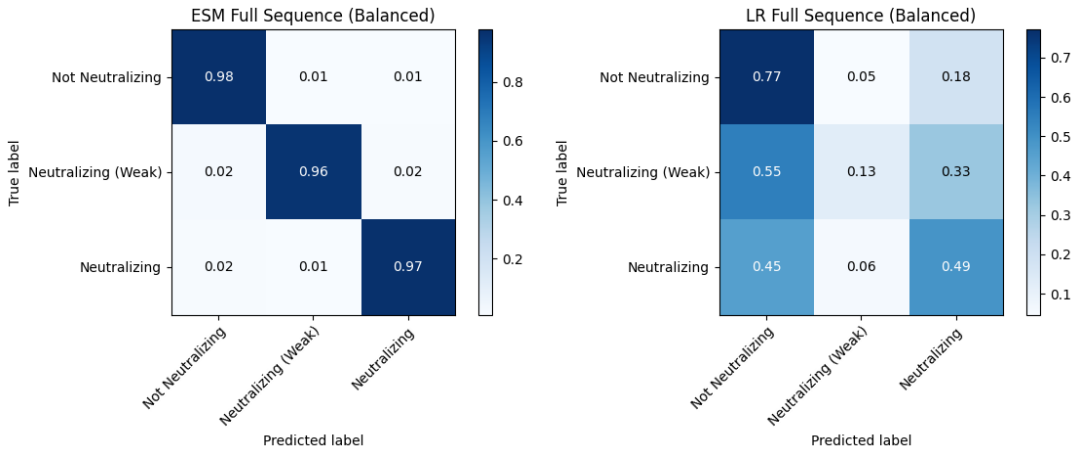

Figure 7: Confusion matrices for the models trained on the balanced dataset, using the full antibody sequences using (a) the ESM model, and (b) Logistic Regression

A Logistic Regression model assumes a linear relationship between the features and log-odds of the target class. Despite proteins being linear polymers of amino acids, the natural folding tendencies of our proteins into complex three-dimensional structures are non-linear. As discussed earlier, these three-dimensional structures are critical for their biological functionality.

Hence, the featurization approach used for the Logistic Regression approach may not necessarily be the best representation of our protein sequences, making traditional ML techniques such as Logistic Regression less effective in the accurate prediction of the neutralization properties between antibodies and antigen sequences using just the protein sequences alone.

When trained on the CDR3 regions of the antibody sequences only, we note that the models generally perform slightly worse than their full antibody sequence counterparts. The true-positive rates drop across all classes in both the transformer-based ESM2 model and the Logistic Regression model, with the only anomaly being the “Not Neutralizing” class in the Logistic Regression model, showing an increase in its true-positive rate from 77% to 79%. This is not surprising, given the Logistic Regression model’s bias, as discussed earlier. The confusion matrices for both models trained on the balanced dataset using the CDR3 regions is shown in Figure 8 below.

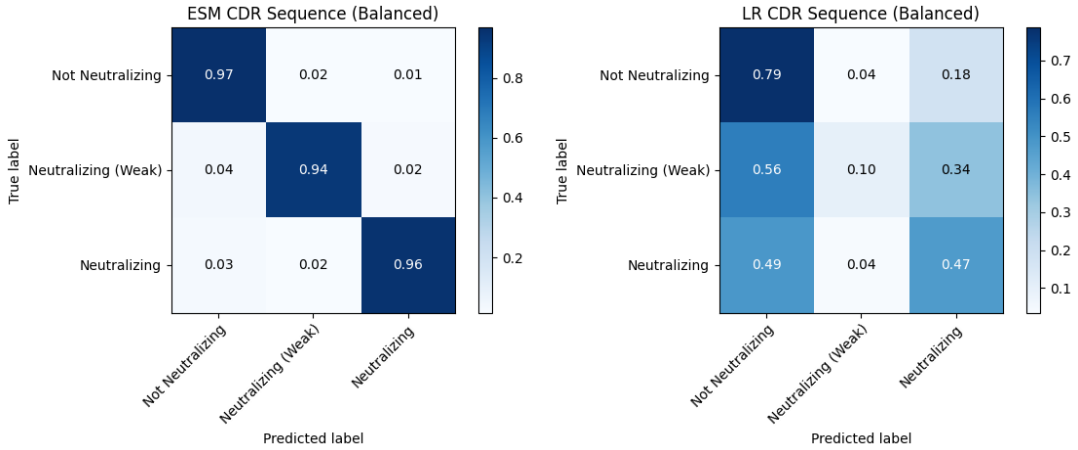

Figure 8: Confusion matrices for the models trained on the balanced dataset, using the CDR3 sequences using (a) the ESM model, and (b) Logistic Regression

CHAPTER 4: FUTURE WORK

4.1 Improvement of dataset accuracy

Given that the CoVAbDb dataset did not provide the original virus sequence, but rather the variant of the virus which each antibody is neutralizing or not neutralizing against, we had to retrieve the referenced protein structures from RSCDB and extract the virus sequences from these protein structures.

These virus sequences were translated into their nucleotide representations using tBLASTn, and put through GISAID's Audacity Instant to determine their variants. Some of these virus sequences were matched to the closest variant(s) defined in the CoVAbDb dataset based on their lineages.

Due to the nature of this mapping, there may be some inaccuracies in the dataset itself, as the generalization of virus sequences across several lineages to their closest variant(s) may affect the overall actual neutralization properties in a real-world scenario.

It is hence recommended the dataset be further refined to attribute the neutralization properties based on specific lineages for the virus sequences, as compared to their WHO-assigned variant names.

4.2 Improvement of dataset accuracy

Despite the labelling of the neutralizing properties within the dataset as "Not Neutralizing", "Neutralizing (Weak)", and "Neutralizing", the neutralization properties of an antibody to a virus is not ternary.

The efficacy of virus neutralization by antibodies is instead determined by metrics such as the half-maximal inhibitory concentration (IC50), and the half-maximal effective concentration (EC50). These metrics represent the concentration of the test substance that is required to neutralize at least 50% of the virus. Hence, a lower IC50 and EC50 value shows signs of a higher efficacy in inhibiting the virus.

Given how the transformer-based protein language model approach has achieved such high accuracies for a ternary classification problem, this method could be further

extrapolated to predict the range of IC50 and EC50 values in a multiclass classification problem, with modifications to the dataset's labels accordingly.

4.3 Use of models that support longer sequences

As the tokenizer for the ESM2 PPLM model could only support sequences of up to 1024 characters, some key examples within the dataset had to be excluded as a result. This includes the examples whereby the SARS-CoV-2 reference protein sequence, the "hCoV-19/Wuhan/WIV04/2019" had to be excluded, as the virus sequence itself was over 1024 characters long. Replicating the experiments using PPLM models that support longer sequences may further validate the findings of this project.

4.4 Adaptation for other Betacoronaviruses

While this project primarily focused on the SARS-CoV2 coronavirus, the CovAbDb dataset records the neutralization properties for other Betacoronaviruses such as the Middle East Respiratory Syndrome (MERS), and SARS.

The methodology could be used to develop similar ML models to predict the neutralization properties with other Betacoronaviruses.

CHAPTER 5: CONCLUSION

In conclusion, we have shown that the use of transformer-based pre-trained protein language models are able to learn and accurately predict the neutralization properties between an antibody and a virus, using just the textual representation of their protein sequences alone, achieving accuracies as high as 98.30%.

We found that the transformer-based PPLM, such as ESM2's ability to learn the relationships between the individual amino acids within a sequence is beneficial, likely due to the self-attention mechanism in transformer-based architectures, as compared to a Logistic Regression model, which may not be the most suitable method, as it assumes a linear relationship between our features.

In addition, we also found that using the full antibody sequence in the training of our models is beneficial, likely due to the additional context the amino acids within the protein sequence outside of the CDR3 regions provide, which is valuable as proteins fold into complex three-dimensional structures, which are critical for their biological functionalities.

The outcomes of this project will be invaluable for the future development of ML models for the efficient validation of potential neutralizing antibodies for viruses.

REFERENCES

[1] Liu, H. *et al.* (2022) 'Human antibodies to SARS-COV-2 with a recurring YYDRxG motif retain binding and neutralization to variants of concern including Omicron', *Communications Biology*, 5(1). doi:10.1038/s42003-022-03700-6.

[2] Pappalardo, F. et al. (2015) "Computational modelling approaches to Vaccinology," Pharmacological Research, 92, pp. 40–45. Available at: https://doi.org/10.1016/j.phrs.2014.08.006.

[3] Magar, R., Yadav, P. & Barati Farimani, A. Potential neutralizing antibodies discovered for novel corona virus using machine learning. Sci Rep 11, 5261 (2021). https://doi.org/10.1038/s41598-021-84637-4.

[4] Ng, Y. H., Chan, J. Y., Wang, C., Kwoh, C. K. and Rashid, S. (2022) "Discovery of Potential Neutralizing Antibodies for SARS-CoV2 through Machine Learning" Proceedings of the URECA@NTU 2021-22. Available at: https://hdl.handle.net/10356/170707.

[5] Huang, Y., Zhang, Z. and Zhou, Y. (2022) "Abagintpre: A deep learning method for predicting antibody-antigen interactions based on sequence information," Frontiers in Immunology, 13. Available at: https://doi.org/10.3389/fimmu.2022.1053617.

[6] Asgari, E. and Mofrad, M.R. (2015) "Continuous distributed representation of biological sequences for deep proteomics and Genomics," PLOS ONE, 10(11). Available at: https://doi.org/10.1371/journal.pone.0141287.

[7] Duan, X. et al. (2022) 'A non-ace2-blocking neutralizing antibody against omicron-included SARS-COV-2 variants', Signal Transduction and Targeted Therapy, 7(1). doi:10.1038/s41392-022-00879-2.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all You need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.

[9] J. Leem, L. S. Mitchell, J. H. R. Farmery, J. Barton, and J. D. Galson, “Deciphering the language of antibodies using self-supervised learning,” Patterns, vol. 3, no. 7, p. 100513, May 2022. https://doi.org/10.1016/j.patter.2022.100513.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, arXiv [cs.CL]. 2019.

[11] D. Wang, F. Ye, and H. Zhou, ‘On pre-trained language models for antibody’, bioRxiv, pp. 2023–2001, 2023.

[12] A. Rives et al., ‘Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences’, Proceedings of the National Academy of Sciences, vol. 118, no. 15, p. e2016239118, 2021.

[13] Y. Liu et al., ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’, arXiv [cs.CL]. 2019.

[14] Matthew I J Raybould and others, CoV-AbDab: the coronavirus antibody database, Bioinformatics, Volume 37, Issue 5, March 2021, Pages 734–735, https://doi.org/10.1093/bioinformatics/btaa739.

[15] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) *Nucleic Acids Research* 28: 235-242 https://doi.org/10.1093/nar/28.1.235.

[16] Peter J. A. Cock and others, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics, Volume 25, Issue 11, June 2009, Pages 1422–1423, https://doi.org/10.1093/bioinformatics/btp163.

[17] Shamima Rashid and others, Jupytope: computational extraction of structural properties of viral epitopes, Briefings in Bioinformatics, Volume 23, Issue 6, November 2022, bbac362, https://doi.org/10.1093/bib/bbac362.

[18] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) “Basic local alignment search tool.” J. Mol. Biol. 215:403-410. PubMed.

[19] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022 Jan 7;50(D1):D20-D26. doi: 10.1093/nar/gkab1112. PMID: 34850941; PMCID: PMC8728269.

[20] Shu, Y. and McCauley, J. (2017) GISAID: from vision to reality. EuroSurveillance, 22(13) doi: 10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101

[21] Ofer, D., Brandes, N. and Linial, M. (2021) 'The language of proteins: NLP, Machine Learning & Protein Sequences', Computational and Structural Biotechnology Journal, 19, pp. 1750–1758. doi:10.1016/j.csbj.2021.03.022.

[22] Z. Lin et al., 'Evolutionary-scale prediction of atomic-level protein structure with a language model', Science, vol. 379, no. 6637, pp. 1123–1130, 2023.

[23] T. Wolf et al., 'HuggingFace's Transformers: State-of-the-art Natural Language Processing', arXiv [cs.CL]. 2020.

APPENDIX A

The source code for this project is available on GitHub:
https://github.com/QixyQix/Antibody-Antigen-Neutralization-Prediction