| | | |
|---|---|---|
| No. | K24-3124 | |
| 研究課題名 | Discovery of antibodies with potential therapeutic applications aided by machine learning and artificial intelligence. | |
| 研究代表者 | Hernandez Jesus ( Centro de Investigación en Alimentación y Desarrollo, A.C.・教授 ) | |
| 研究組織 | 受入教員 | 中井 謙太 ( 東京大学医科学研究所・教授 ) |
| | 分担者 | MARTIN DE JESUS LOZA LOPEZ ( Laboratory of Functional Analysis in silico・Assistant Professor ) |
| | 分担者 | DIANA GABRIELA HINOJOSA TRUJILLO ( Centro de Investigacion en Alimentacion y Desarrollo, AC/ Laboratory of Immunology・Doctoral student ) |
| | 分担者 | MONICA RESENDIZ SANDOVAL ( Centro de Investigacion en Alimentacion y Desarrollo, AC/ Laboratory of Immunology・Associate Professor ) |
| | 分担者 | ANA MELISSA GARCIA VEGA ( Centro de Investigacion en Alimentacion y Desarrollo, AC/ Laboratory of Immunology・Doctoral student ) |

(Form 2)

| Principal Investigator |
| --- |
| Prof. Jesus Hernandez |

**IMSUT International Joint Usage/Research Center Project <International>**

**Joint Research Report (Annual/Project Completion)**

**Annual Report**

## Report

During this research period we curated a dataset comprising antibody sequences targeting SARS-CoV and SARS-CoV-2. Specifically, we collected the heavy and light chain sequences of antibodies, along with experimentally measured IC50. (Half-maximal inhibitory concentration) values indicating their neutralization potency. Approximately 7,000 antibodies with IC50 measurements against the wild-type (WT) strain of SARS-CoV-2 were included, along with around 3,500 IC50 values for major Omicron variants (BA.1, BA.2, BA.2.75, BA.5, and XBB). Data collection was performed by mining The Coronavirus Antibody Database (CoV-AbDab) database created by The Oxford Protein Informatics Group (Department of Statistics, University of Oxford) to retrieve antibody sequences against SARS-CoV-2. Additionally, we searched the original publications to obtain the experimental IC50 (half maximal inhibitory concentration) values derived from pseudovirus and live-virus neutralization assays which represent the concentration of an antibody needed to inhibit 50% of viral activity, serving as a key measure of antibody potency in neutralization assays. In total, 350 publications were reviewed to compile the data.

For predicting antibody neutralization, we proposed leveraging AntiBERTy, a pretrained BERT-based model specifically optimized for antibody sequence analysis (Figure 1). AntiBERTy utilizes the rich semantic features captured from antibody sequences to predict their neutralization potential based on sequence characteristics. Built on a standard BERT architecture, AntiBERTy has an embedding size of 512 and is adapted to handle a specialized vocabulary of 25 tokens, allowing it to effectively encode antibody-related inputs, such as amino acid sequences. The model incorporates positional and token-type embeddings, which will later be utilized in unsupervised analysis. Its encoder is comprised of 8 transformer layers, each featuring self-attention mechanisms and feed-forward networks. In addition to the core BERT architecture, AntiBERTy includes specialized prediction heads designed for multiple tasks: sequence prediction (language modeling), species classification, chain type identification (heavy or light chain), and grafting site prediction, all utilizing linear layers to address these specific challenges in antibody analysis.

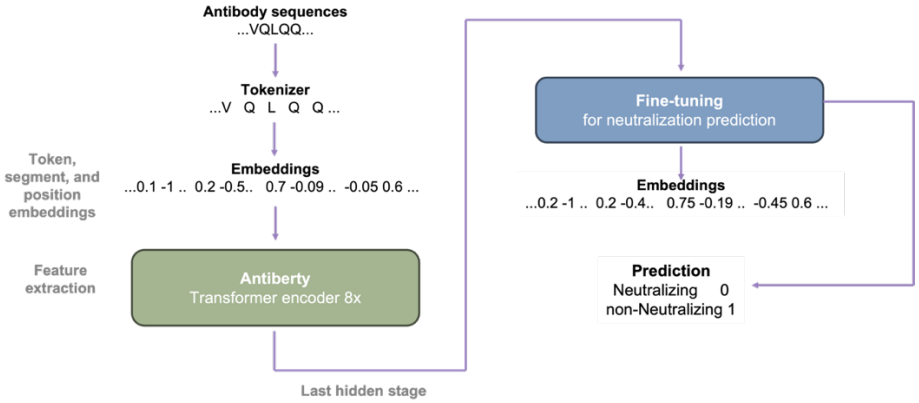


*Figure 1. Overview of the antibody neutralization prediction framework.*

Building on this foundation, we extend the pretrained AntiBERTy model by introducing a new module specifically designed for predicting antibody neutralization capacity, allowing it to directly associate sequence features with functional neutralization outcomes against specific antigens, in this case SARS-CoV-2. To extend the capabilities of AntiBERTy beyond sequence representation, we introduced a neutralization prediction head designed to classify antibody sequences based on their potential neutralizing activity. This additive module operates on top of the pretrained AntiBERTy encoder, utilizing the [CLS] token embedding from the final hidden layer as input. The main idea is to leverage the rich semantic features captured by AntiBERTy and adapt them to a supervised neutralization task, allowing fine-grained discrimination of antibody function without retraining the entire backbone.

Several multilayer perceptron (MLP) architectures are being systematically evaluated to optimize prediction performance.

These architectures vary in depth, hidden dimensionality, and activation functions. Simpler designs include a single linear layer directly projecting the 512-dimensional [CLS] embedding to a scalar output, while more complex configurations incorporate one to four hidden layers with dimensions ranging from 32 to 256 neurons. Activation functions such as ReLU, LeakyReLU, and SELU are tested in combination with regularization strategies including dropout or alpha dropout, and in some cases batch normalization. In parallel, we are testing multiple representations of antibody sequences, including using the heavy chain or light chain individually, combining both chains, and creating sequences from the complementarity-determining regions (CDRs) to identify the most effective input format for neutralization prediction.

| Model number | Antibody sequence | AUC | F1 score | Recall | Precision | Accuracy |
|---|---|---|---|---|---|---|
| 11 | heavy chain | 0.701 | 0.716 | 0.648 | 0.801 | 0.690 |
| 9 | H1L1 | 0.702 | 0.739 | 0.698 | 0.784 | 0.701 |
| 3 | heavy chain | 0.704 | 0.742 | 0.704 | 0.785 | 0.704 |
| 10 | H1L1 | 0.705 | 0.750 | 0.723 | 0.779 | 0.709 |
| 12 | L1H1 | 0.705 | 0.811 | 0.765 | 0.864 | 0.735 |
| 0 | H1L1 | 0.706 | 0.718 | 0.648 | 0.809 | 0.693 |
| 10 | heavy chain | 0.706 | 0.729 | 0.669 | 0.800 | 0.699 |
| 0 | H1L1 | 0.707 | 0.711 | 0.630 | 0.817 | 0.691 |
| 3 | H1L1 | 0.712 | 0.738 | 0.684 | 0.801 | 0.706 |
| 2 | heavy chain | 0.712 | 0.731 | 0.667 | 0.808 | 0.703 |
| 9 | H1L1 | 0.713 | 0.715 | 0.631 | 0.825 | 0.696 |
| 8 | heavy chain | 0.723 | 0.719 | 0.629 | 0.840 | 0.703 |

*Table 1. Performance metrics across different model configurations for antibody neutralization prediction. Several model architectures were evaluated using different antibody sequence inputs.*

So far, the results are promising. As shown in the Table 1, several model configurations achieved strong predictive performance across multiple evaluation metrics. Specifically, the best-performing models demonstrated high AUC (Area Under the Curve) values and strong F1 scores in the validation data. Importantly, we observed that the best-performing models consistently utilized the same antibody sequence representation, highlighting that the choice of input encoding has a major impact on prediction performance. In particular, using only the heavy chains and combining the CDRs (complementarity-determining regions) of both heavy and light chain sequences in particular orders, provided a more comprehensive characterization of the antibody and improved the model's ability to capture relevant functional features associated with neutralization. Preliminary analysis of the embeddings shows grouping of highly neutralizing antibodies (yellow points) particularly in the circled region (Figure 2). This suggests that fine-tuning helps to subtly refine the antibody representations, making functional patterns more apparent. These preliminary results are encouraging and demonstrate that AntiBERTy-based embeddings, when fine-tuned with a dedicated neutralization prediction head, can accurately capture functional antibody properties directly from sequence information.

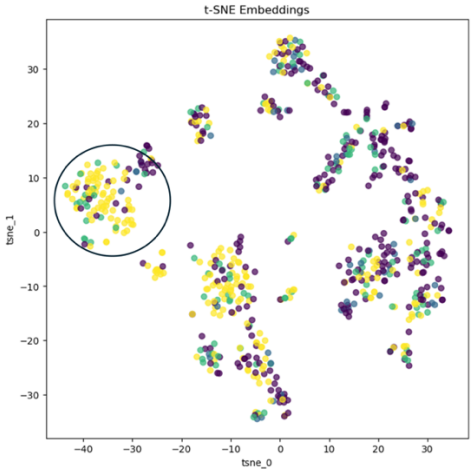


*Figure 2. t-SNE visualization of antibody sequence embeddings after fine-tuning. Colors represent antibody neutralization potency based on $IC_{50}$ values (purple: non-neutralizing, green: moderate neutralization, blue: low neutralization, yellow: high neutralization).*

In summary, we have successfully collected the data and tested neutralization predictions using AntiBERTy with different predictor layers. We have also visualized the AntiBERTy embeddings of antibodies. In further steps, we will try to improve the prediction metrics and analized the embeddings using unsupervised learning to select antibodies with broadly neutralizing capabilities. We expect to experimentally produce the antibodies at the end of the fiscal year 2025.