No.	25-1154	
Project Title	Integrated machine learning and large language models to identify broadly neutralizing antibodies with therapeutic potential for viral diseases caused by the Flavivirus group	
Principal Investigator	Kenji Mizuguchi (Institute for Protein Research, Osaka University・Professor)	
Project Member(s)	IMSUT Host Researcher	Kenta Nakai (The Institute OF Medical Science, The University Of Tokyo・Professor)
	Project Member (s)	LOZA LOPEZ MARTIN DE JESUS (The laboratory of Functional Analysis in-silico・Assistant Professor)
	Project Member (s)	NATSRITA PIYATIDA (Institute for Protein Research, Osaka University・Doctoral student)

(Form 2)

Principal Investigator
MIZUGUCHI KENJI

IMSUT International Joint Usage/Research Center Project <Domestic>

Joint Research Report (Annual/Project Completion)

Annual Report

Report

Research Summary

This project aims to identify broadly neutralizing antibodies (bNAbs) for treating flavivirus infections, which include Zika (ZIKV), Dengue (DENV), Japanese encephalitis (JEV), and West Nile (WNV). These viruses cause around 300 million infections annually, with 50–100 million symptomatic cases. Effective treatment options remain limited, mainly due to antibody-dependent enhancement (ADE), where non-neutralizing antibodies can worsen disease outcomes. To overcome this, we proposed an integrated approach using machine learning (ML) and large language models (LLMs) to rapidly identify bNAbs across multiple flavivirus serotypes. The project leverages sequence and experimental data, aiming to accelerate antibody discovery and reduce development costs.

Progress and Achievements – FY2025

The work completed this year aligns with the main steps outlined in the grant application, particularly focusing on data integration, model development (Figure 1), and initial candidate screening. We have successfully completed the following tasks:

- *Database construction* (Step 1 in the original proposal):

  We established the Flavi-Ab database by collecting and integrating antibody sequence information from 296 publications and 22 patent documents. The database now includes 3,984 monoclonal antibody entries targeting 16 flaviviruses, with 3,448 CDR-H3 sequences, 6,541 epitope records, and 7,550 experimentally validated neutralization records.

- *Feature extraction and model architecture* (Step 2 in the original proposal; Figure 1):

  A comprehensive feature extraction strategy was implemented, using 31 descriptors from three main types:
  - Sequence-based descriptors, capturing amino acid composition, arrangement patterns, and physicochemical properties of CDR-H3 sequences.
  - Fingerprint-based descriptors, transforming CDR-H3 sequences into fixed-length vectors that encode local residue arrangements and motif-like features. These descriptors are designed to capture pattern-level information that is not fully addressed by global sequence-based metrics.
  - Language model-based descriptors, from pretrained models such as AntiBERTa2, were used to extract contextualized embeddings from antibody sequences. These embeddings encode structural, functional, and evolutionary information relevant for antibody–antigen interaction prediction.

- *Performance comparison and ensemble-model development* (Step 2 in the original proposal):

  We constructed 186 baseline models by training six machine-learning algorithms (Extra Trees, Logistic Regression, Multilayer Perceptron, Random Forest, Support Vector Machine, and XGBoost) with the 31 features described above. Model performance was

assessed using five-fold cross-validation and independent test sets, with metrics including accuracy, sensitivity, specificity, Matthews correlation coefficient, and AUC. The top 10 models ranked by AUC were selected as informative features for cross-neutralization prediction as shown in Figure 1.

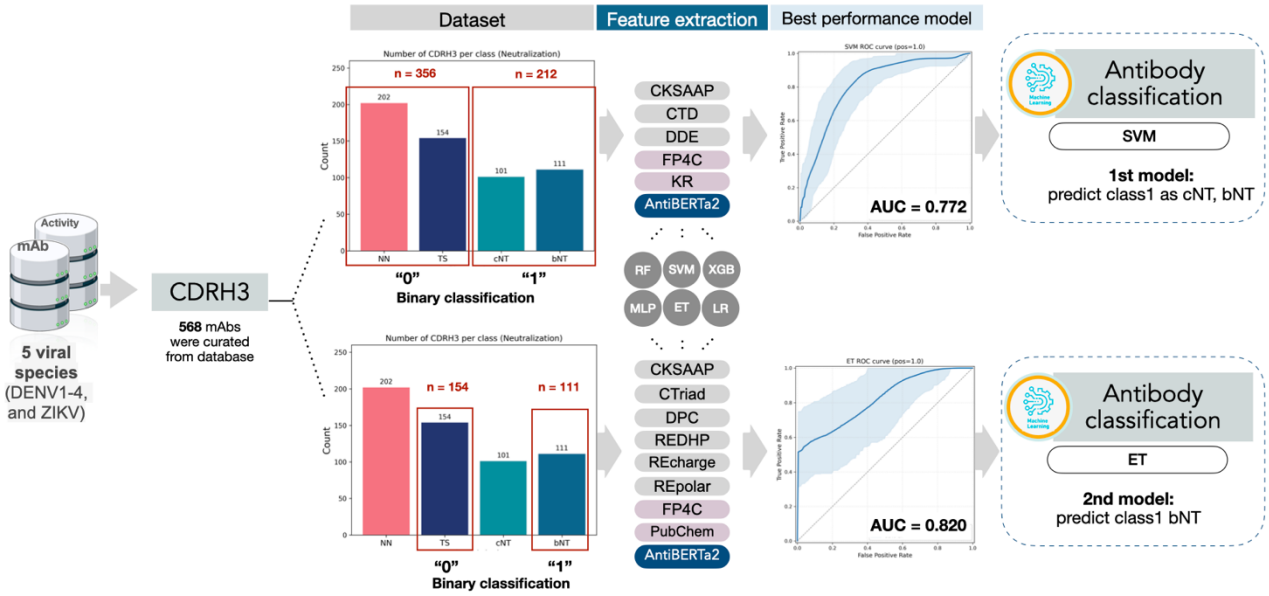

Figure 1 LLM-integrated model construction workflow

- Artificial Library Generation (Step 4):

  Although this was originally scheduled for the second year, we initiated the generation of an in-silico library comprising 2,363,095 CDR-H3 variants. This library was created by systematically introducing single, double, and triple amino acid substitutions into the CDR-H3 loop of the weakly neutralizing human monoclonal antibody 2D1G5 (IC50 > 10 μg/mL for all four DENV serotypes). Variants were screened using a two-stage computational pipeline (Figure 2): first, ranking by two high-performing machine-learning models, and second, assessing stability and heavy–light chain compatibility using ImmunoMatch. This approach enables high-throughput identification of novel broadly neutralizing antibody candidates.

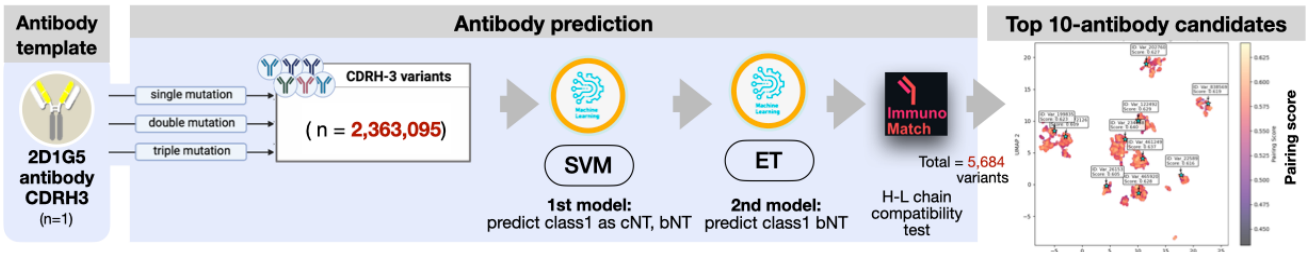

Figure 2 LLM-integrated model construction workflow

Plans for FY2026

In accordance with the grant application, the next steps will focus on candidate characterization and experimental validation:

- Structural Modeling and Docking:

  De novo modeling of CDR-H3 regions and VH–VL orientation is ongoing using RosettaAntibody3. Molecular docking will be performed against DENV-1–4 and ZIKV envelope proteins to assess binding interactions using ZDOCK server. Based on these results, a subset of approximately 10 antibody candidates will be selected for experimental validation of antigen recognition and neutralization.

- As an additional computational analysis, we are generating in silico CDR-H3 variant libraries for two other antibodies, 2D22 and ZIKV-195, with narrow specificity profiles. 2D1G5

antibody neutralizes only DENV-3 but recognizes several (DENV1-4), while the other both recognizes and neutralizes only one variant (DENV-2 and ZIKV, respectively). The goal is to enhance their breadth of recognition and neutralization through systematic CDR-H3 modifications.

- We are also preparing a manuscript to report our findings.

Collaboration and Research Environment

The collaboration between The University of Osaka and the University of Tokyo is progressing well. We hold regular online meetings to review results, discuss ongoing work, and define new tasks. These meetings have been valuable for project coordination and troubleshooting. In addition, several in-person visits between the institutes have taken place, further strengthening the collaboration and providing opportunities for knowledge exchange.