No.	K24-3124	
Project Title	Discovery of antibodies with potential therapeutic applications aided by machine learning and artificial intelligence.	
Principal Investigator	Jesus Hernandez (Centro de Investigación en Alimentación y Desarrollo, A.C.・Professor)	
Project Member(s)	**IMSUT Host Researcher**	Kenta Nakai (The Institute OF Medical Science, The University Of Tokyo・Professor)
	Project Member (s)	MARTIN DE JESUS LOZA LOPEZ (Laboratory of Functional Analysis in silico・Assistant Professor)
	Project Member (s)	DIANA GABRIELA HINOJOSA TRUJILLO (Centro de Investigacion en Alimentacion y Desarrollo, AC/ Laboratory of Immunology・Doctoral student)
	Project Member (s)	MONICA RESENDIZ SANDOVAL (Centro de Investigacion en Alimentacion y Desarrollo, AC/ Laboratory of Immunology・Associate Professor)

(Form 2)

Principal Investigator
Prof. Jesus Hernandez

IMSUT International Joint Usage/Research Center Project <International>

Joint Research Report (Annual/Project Completion)

Project Completion Report

Report

During this research period, we curated a comprehensive dataset comprising antibody sequences targeting 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) to retrieve antibody sequences against SARS-CoV-2. Additionally, we searched the original publications to obtain experimental IC50 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 total, 350 publications were reviewed to compile this data.

For predicting antibody neutralization, we leveraged AntiBERTy, a pretrained BERT-based model specifically optimized for antibody sequence analysis (Figure 1).The main idea is to leverage the rich semantic features captured by AntiBERTy and adapt them to a supervised neutralization task, allowing the identification of antibodies with neutralization capability.

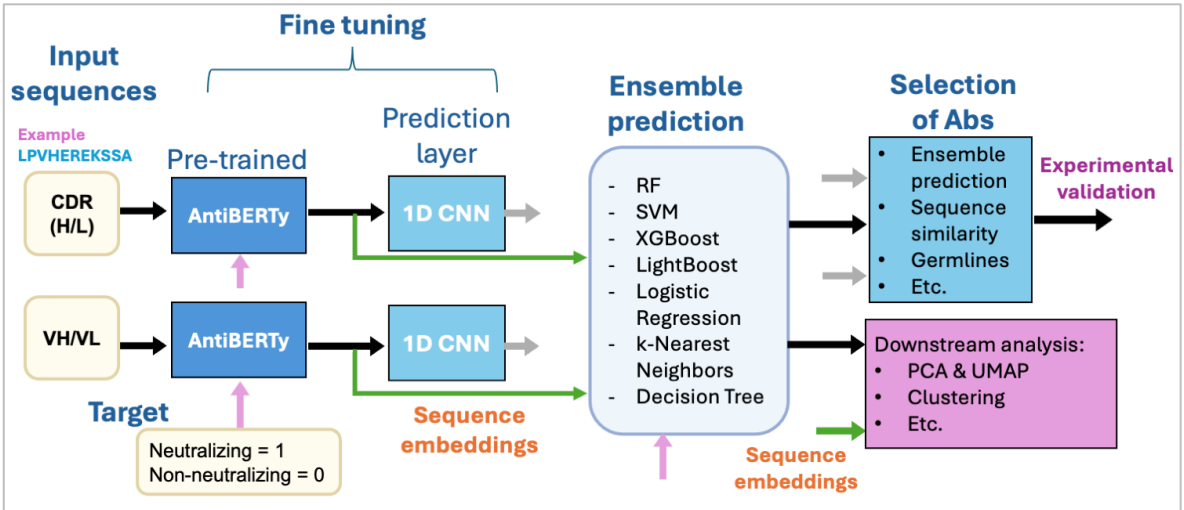

Figure 1. Workflow analysis of antibody neutralization prediction, downstream analysis and selection of final set of antibodies for experimental validation

We fine-tuned the pretrained AntiBERTy model by introducing a new module specifically designed to predict antibody neutralization capacity. To optimize feature representation, we tested two approaches for encoding antibody sequences: whole chain and CDR-only representations. The embeddings from both representations were then merged and used as input features for an ensemble prediction framework comprising multiple machine learning classifiers, including random forest and support vector machine models. The final predictions were determined by selecting antibodies that achieved the highest scores across all methods.

So far, the results are promising. As shown in 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 approaching 0.79 and strong F1 scores in the

validation data. Importantly, we observed that the best-performing models consistently utilized similar antibody sequence representations, 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 clear grouping of highly neutralizing antibodies particularly in distinct clusters (Figure 2). This suggests that fine-tuning helps to subtly refine the antibody representations, making functional patterns more apparent. The embeddings also showed strong correlation with biological features such as germline gene usage (Figure 3), confirming that the model captures biologically meaningful patterns directly from sequence information.

To extend our computational predictions, we applied the trained models to a private dataset of 22,915 antibody sequences derived from SARS-CoV-2 RBD/S1-positive B cells from COVID-19 convalescent patients. Using a multi-stage selection strategy, we first identified 378 candidates based on ensemble predictions from both CNN and machine learning models. These candidates were then filtered based on sequence similarity to known broadly neutralizing antibodies (>0.80 similarity in both heavy and light chains), reducing the pool to 62 antibodies. Final selection criteria included germline genes linked to neutralizing responses, neutralization-associated motifs, high similarity to Omicron-neutralizing antibodies, and antibody isotype preferences (predominantly IgG1 and selected IgM), resulting in 23 antibodies selected for experimental validation.

	Precision	Recall	F1	MCC	AUC ROC
CNN CDRs	0.7633	0.7726	0.7679	0.4251	0.7661
CNN VH/VL	0.7163	0.8704	0.7858	0.4035	0.7899
CDR embeddings					
LGBM	0.7436	0.7873	0.7648	0.3986	0.7535
SVM	0.7354	0.8019	0.7672	0.3926	0.7499
XGB	0.7291	0.8093	0.7671	0.3848	0.7626
KNN	0.7406	0.7751	0.7575	0.3847	0.7386
RF	0.7275	0.8093	0.7662	0.3815	0.7688
VH/VL embeddings					
RF	0.7289	0.7824	0.7547	0.3665	0.7695
LGBM	0.7375	0.7555	0.7464	0.3662	0.7716
KNN	0.7356	0.7482	0.7418	0.3582	0.7319
XGB	0.7264	0.7726	0.7488	0.3554	0.7715
SVM	0.7130	0.7653	0.7382	0.3236	0.7264

Table 1. Prediction results' metrics per method

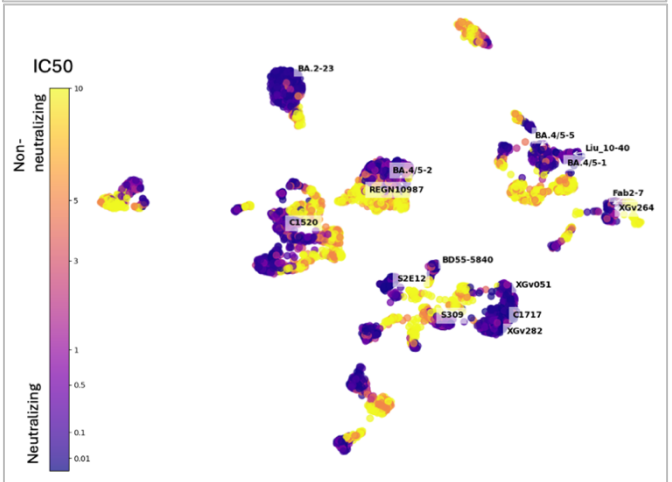

Figure 2. UMAP representation of AntiBERTy embeddings shows clear distinction of neutralizing and non neutralizing antibodies

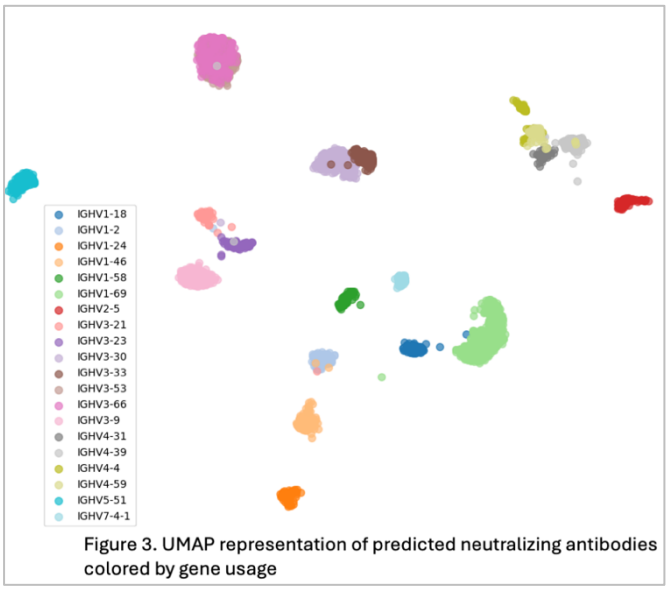

Figure 3. UMAP representation of predicted neutralizing antibodies colored by gene usage

Initial binding screening results have been highly encouraging. Among the 23 tested antibodies, most showed binding recognition to the wild-type (WT) SARS-CoV-2 variant, which was the variant used for model training; only five candidates did not show detectable recognition. Notably, three antibodies exhibited broad binding profiles, recognizing all variants included in the ELISA screening panel (Figure 4). These candidates are currently undergoing further experimental validation, including larger-scale recombinant expression to obtain sufficient material for detailed binding characterization against the receptor-binding domain (RBD) of WT and emerging variants, including recent lineages such as JN.1. Comprehensive pseudovirus neutralization assays are also planned in collaboration with Karolinska Institute. In parallel, molecular docking studies are being conducted with collaborators at Osaka University to define binding epitopes and better understand the interaction mechanisms of these promising candidates.

In summary, we have successfully developed a machine learning pipeline that identifies functional neutralizing antibodies directly from sequence data. Our AntiBERTy-based models, trained on a comprehensive curated dataset, correctly predicted 3 broadly neutralizing candidates out of 23 tested antibodies - demonstrating proof-of-concept for computational antibody discovery. Full experimental characterization is expected by fiscal year-end 2026, with manuscript preparation underway.

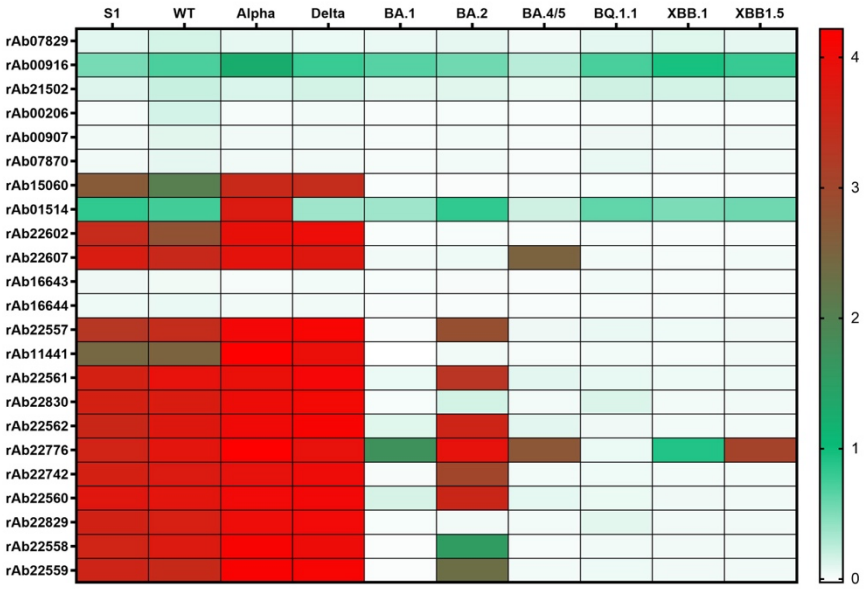

Figure 4. Heatmap of preliminary ELISA binding recognition of the 23 antibodies against SARS-CoV-2 variants.