A SARS-CoV-2 Interaction Dataset and VHH Sequence Corpus for Antibody Language Models

Hirofumi Tsuruta[1,2], **Hiroyuki Yamazaki**[1,3], **Ryota Maeda**[1,3],
Ryotaro Tamura[1,2], **Akihiro Imura**[1,3]
[1]COGNANO Inc., [2]SAKURA internet Inc., [3]Biorhodes, Inc.
{tsuruta, yamazaki, maeda, ryotarotamura, akihiroimura}@cognano.co.jp

Abstract

Antibodies are crucial proteins produced by the immune system to eliminate harmful foreign substances and have become pivotal therapeutic agents for treating human diseases. To accelerate the discovery of antibody therapeutics, there is growing interest in constructing language models using antibody sequences. However, the applicability of pre-trained language models for antibody discovery has not been thoroughly evaluated due to the scarcity of labeled datasets. To overcome these limitations, we introduce AVIDa-SARS-CoV-2, a dataset featuring the antigen-variable domain of heavy chain of heavy chain antibody (VHH) interactions obtained from two alpacas immunized with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike proteins. AVIDa-SARS-CoV-2 includes binary labels indicating the binding or non-binding of diverse VHH sequences to 12 SARS-CoV-2 mutants, such as the Delta and Omicron variants. Furthermore, we release VHHCorpus-2M, a pre-training dataset for antibody language models, containing over two million VHH sequences. We report benchmark results for predicting SARS-CoV-2-VHH binding using VHHBERT pre-trained on VHHCorpus-2M and existing general protein and antibody-specific pre-trained language models. These results confirm that AVIDa-SARS-CoV-2 provides valuable benchmarks for evaluating the representation capabilities of antibody language models for binding prediction, thereby facilitating the development of AI-driven antibody discovery. The datasets are available at `https://datasets.cognanous.com`.

1 Introduction

Antibodies are vital proteins produced by the immune system to remove harmful foreign substances called antigens. Antibody-based therapeutics, which can bind to target antigens with high affinity and specificity, have become a major class of therapeutic agents and are currently used to treat a wide range of diseases [1, 2]. Among their successes, the rapid development and subsequent approval of antibodies against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epitomize the impactful response of this therapeutic class in addressing urgent global health challenges [74, 30]. However, the development of therapeutic antibodies remains a time-consuming and costly endeavor due to the complexity and difficulty of artificially manipulating the vast search space of antibody sequences [21]. Therefore, computational approaches for accelerating antibody discovery have become increasingly popular in recent years [72, 24, 4].

Recent advances in language models offer new possibilities for understanding the information contained in antibody sequences because an antibody sequence can be represented as a string of letters representing a type of amino acid. With the construction of the observed antibody space (OAS) database [26, 46] that currently contains over two billion antibody sequences, a sufficient number of antibody sequences is now available to train antibody-specific language models [56, 27, 47, 70, 49, 5,

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-3688

22]. Olsen *et al.* [47] presented AbLang, an antibody language model pre-trained on either the heavy or light chain antibody sequences in the OAS database. They demonstrated that AbLang can be used to accurately restore the missing residues in antibody sequences. Wang *et al.* [70] proposed EATLM, a pre-trained antibody language model that incorporates evolutionary information as the pre-training objectives. They also provided an antibody understanding evaluation (ATUE) benchmark consisting of four tasks to evaluate the performance of pre-trained language models in antibody-related tasks.

Despite these promising developments, the applicability of pre-trained language models for antibody discovery has not been adequately evaluated due to the lack of labeled datasets. ATUE includes an antibody discovery task, a binary sequence classification that distinguishes antibodies that bind to SARS-CoV-2. The training dataset for this task used antibody sequences from SARS-CoV-2 patients and healthy persons from the OAS database. Although very few antibodies from SARS-CoV-2 patients are directly responsible for virus binding, these noisy and potentially unreliable individual-level disease labels were used to train a sequence-level classifier. Thus, a dataset with labels indicating whether the antibody binds to a specific antigen at the antibody sequence level would be extremely useful for a more accurate evaluation of model performance for antibody discovery.

In this study, we introduce AVIDa-SARS-CoV-2, a dataset featuring the antigen-variable domain of heavy chain of heavy chain antibody (VHH) interactions produced by two alpacas immunized with SARS-CoV-2 spike proteins. VHHs, found in camelids such as alpacas and llamas, are promising therapeutic agents because of their small size, high stability, and high antigen-binding affinity [20, 19]. AVIDa-SARS-CoV-2 was generated using our previously established method for generating interaction datasets with reliable labels [66]. AVIDa-SARS-CoV-2 contains binary labels that indicate whether each of the diverse VHH sequences binds or does not bind to 12 SARS-CoV-2 mutants, such as the Delta and Omicron variants. Notably, label reliability was verified by experimental evidence that VHHs extracted from AVIDa-SARS-CoV-2 bound to SARS-CoV-2 spike variants [38].

Furthermore, we introduce VHHCorpus-2M, a pre-training dataset for antibody language models containing over two million VHH sequences, and VHHBERT, an antibody language model pre-trained on VHHCorpus-2M. To avoid sequencing errors and increase sequence reliability, we removed singletons from the VHH sequences identified by next-generation sequencing (NGS), that is, only sequences observed more than once were used in the corpus. Although VHHCorpus-2M contains fewer sequences than OAS, it is distinctive in that it consists entirely of full-length VHH sequences that act as the smallest functional units for binding to each target antigen.

The main contributions of this paper are summarized as follows.

- We release AVIDa-SARS-CoV-2, a labeled SARS-CoV-2-VHH interaction dataset with amino acid sequences, and VHHCorpus-2M, which contains over two million unlabeled VHH sequences. These datasets can be used for the evaluation and pre-training of antibody-specific language models.
- AVIDa-SARS-CoV-2 contains information on the interactions of diverse VHHs produced by two alpacas with 12 SARS-CoV-2 mutants, providing researchers with valuable insights into the effects of antigen mutations on antibody binding and individual differences in antigen-specific VHHs.
- We release VHHBERT, a VHH-specific language model pre-trained using VHHCorpus-2M. VHHBERT will serve as a baseline for subsequent VHH-specific language models.
- We report benchmark results for the prediction of SARS-CoV-2-VHH interactions using VHHBERT and existing general protein and antibody-specific pre-trained language models. These results confirm that AVIDa-SARS-CoV-2 provides valuable benchmarks for assessing the representation capabilities of antibody language models for binding prediction.

2 Related Work

In this section, we put our work in the context of existing pre-trained antibody language models and their datasets used for pre-training and evaluation. Currently, there is growing interest in constructing language models using protein sequences [55, 52, 15, 32]. Inspired by these successes and the fact that the evolutionary process of antibodies is significantly different from that of proteins, several studies have attempted to train language models specific to antibody sequences. The representative existing studies are summarized in Table 1.

https://doi.org/10.52202/079017-3688

Table 1: Characteristics of pre-trained antibody language models. "M" stands for million, and "B" stands for billion.

	Pre-training			Evaluation		
Model	Dataset	#Samples	Chain Type	Dataset	#Samples	Task
AntiBERTy [56]	OAS	588M	Heavy, light	HIV-1 donor repertoires [76, 75]	232,593	Evolutionary analysis
AntiBERTa [27]	OAS	72M	Heavy, light	SAbDab [14]	900	Paratope prediction
AbLang-H [47]	OAS	14M	Heavy	OAS	2,000	Sequence restoration
AbLang-L [47]	OAS	0.24M	Light	OAS	4,200	Sequence restoration
EATLM [70]	OAS	20M	Heavy, light	Mason *et al.*'s dataset [40]	21,612	Binding prediction
				SAbDab [14]	1,662	Paratope prediction
				Mroczek *et al.*'s dataset [43]	88,094	B cell classification
				OAS, CoV-AbDab [53]	22,000	Antibody discovery
BERT-DS [49]	OAS	20M	Heavy	HER2affmat	234,088	Binding prediction
AntiBERTa2 [5]	OAS, proprietary dataset	824M	Heavy, light	Mason *et al.*'s dataset [40]	22,779	Binding prediction
IgBert [22]	OAS	2B	Heavy, light	OAS	20,000	Sequence restoration
				FLAb [10]	6,745	Binding affinity prediction
				OAS	1,000	Perplexity
VHHBERT	**VHHCorpus-2M**	**2M**	**Heavy**	**AVIDa-SARS-CoV-2**	**77,003**	**Binding prediction**

Pre-trained Antibody Language Models. Ruffolo *et al.* [56] proposed AntiBERTy, the first antibody-specific language model to understand affinity maturation within immune repertoires. They found that AntiBERTy can cluster antibodies into trajectories resembling affinity maturation. Leem *et al.* [27] presented a pre-trained antibody language model called AntiBERTa and fine-tuned AntiBERTa to predict the binding site of an antibody, called a paratope, from an antibody sequence. Olsen *et al.* [47] pre-trained two language models: one trained only on heavy chains of antibodies (AbLang-H) and one trained only on light chains of antibodies (AbLang-L). These models outperformed ESM-1b [55], a general protein language model, in restoring the missing residues in antibody sequences. Wang *et al.* [70] incorporated two original pre-training objectives into their proposed antibody language model, EATLM, to explore the benefits of incorporating specific biological mechanisms into pre-training. They also provided a useful benchmark, ATUE, that consists of four antibody-related tasks. Porebski *et al.* [49] pre-trained BERT-DS using 20 million heavy-chain human antibody sequences and fine-tuned it for binding prediction. Barton *et al.* [5] developed AntiBERTa2, an antibody language model pre-trained using 824 million antibody sequences including paired heavy and light chain sequences, and proposed a multimodal contrastive learning that amalgamates the representations of antibody sequences and structures. Kenlay *et al.* [22] presented IgBert, which was initialized with the pre-trained protein language model ProtBERT [15] and trained using more than two billion unpaired antibody sequences (i.e., heavy chain or light chain only) and two million paired antibody sequences. Various other antibody-specific language models have been developed for antibody humanization [50], sequence generation [61, 45, 41], identification of evolutionarily plausible mutations [18], and classification of antigen-specific antibodies [8].

Pre-training Datasets. The pre-training datasets for antibody language models are large collections of unlabeled antibody sequences. Conventional antibodies in humans and mice comprise two pairs of heavy and light chains, meaning that one pair of chains serves as the functional unit for binding to the target antigen. Currently, OAS [26, 46] contains over two billion unpaired antibody sequences, more than 90% of which are of human origin. Existing antibody language models are pre-trained primarily using unpaired antibody sequences in the OAS database. AntiBERTa2 and IgBert were trained with paired antibody sequences, but with a very small proportion compared with unpaired sequences.

VHHs are variable regions of heavy-chain antibodies found in camelids. Because VHH acts as a single functional unit, its sequence contains all the information necessary for antibody functions against antigens. The OAS database currently contains approximately 1.6 million unique VHH sequences collected from one study [31] and derived from three unimmunized Bactrian camels. The Integrated Nanobody Database for Immunoinformatics (INDI) database [12] currently contains more than 11 million unique VHH sequences collected mainly from the Sequence Read Archive (SRA) [28] and derived from dromedaries, Bactrian camels, llamas, and alpacas. VHHCorpus-2M contains more than two million unique VHH sequences generated in original experiments using five alpacas.

Evaluation Datasets. The evaluation datasets are a set of labeled antibody sequences used to assess model performance for a specific task. AntiBERTa and EATLM were fine-tuned using the structural antibody database (SAbDab) [14] to evaluate their performance in predicting paratopes. Paratope prediction is important for the efficient discovery of antibody candidates that bind to an

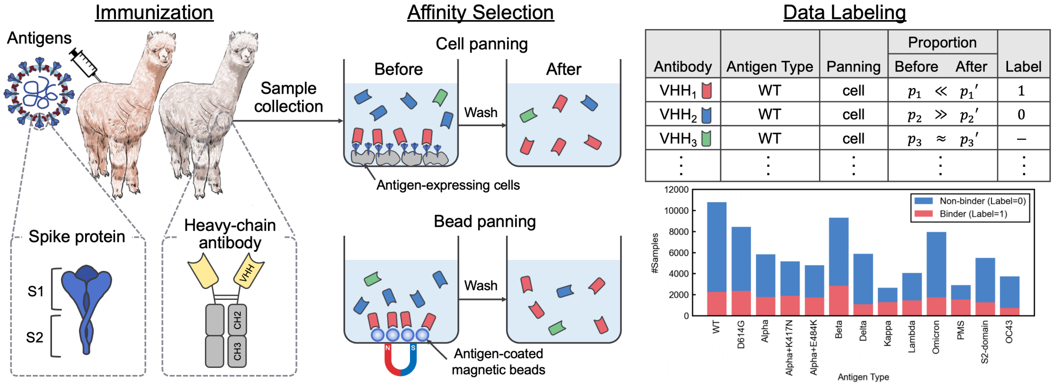

Figure 1: Overview of data generation process for AVIDa-SARS-CoV-2.

antigen of interest; however, the size of labeled datasets is limited. IgBert was evaluated in a binding affinity prediction task using each of three datasets with small data samples of 422 [59], 2048 [71], and 4275 [25] from the fitness landscape for antibodies (FLAb) [10]. BERT-DS was evaluated in a binding prediction task involving a three-category classification using a deep-screening dataset called HER2affmat. Although HER2affmat is a useful dataset with a large number of samples, it does not contain full-length antibody sequences. A binding prediction task using Mason *et al.*'s dataset [40] in ATUE [70] was done that involved binary classification to determine whether the complementarity-determining region (CDR) of an antibody can bind to human epidermal growth factor receptor 2 (HER2). AntiBERTa2 was also evaluated for its performance in binding prediction using the same dataset. All antibody sequences in this dataset were derived from a single germline sequence, indicating that the diversity of antibody sequences was strongly limited. Thus, in ATUE, this task is considered to be less relevant to antibody-specific evolution.

The antibody discovery task in ATUE is a binary sequence classification that distinguishes antibodies that bind to SARS-CoV-2. This task has two notable limitations in terms of accurate model evaluation for antibody discovery. First, the dataset for training the sequence classifier uses antibody sequences with noisy individual-level labels from SARS-CoV-2 patients and healthy persons, even though very few antibodies from SARS-CoV-2 patients are responsible for virus binding. Second, this task assumes that if the third CDR of the heavy chain (CDR-H3) of the binder sequence predicted by the model is 90% or more identical to the CDR-H3 of the true binding sequences in CoV-AbDab [53], they have a similar binding performance. However, not only the sequence of CDRs but also the appropriate three-dimensional structure and interactions between variable regions are important for antigen-antibody activity [44]. AVIDa-SARS-CoV-2 has sequence-level labels for binding and non-binding to SARS-CoV-2 mutants for each full-length VHH sequence.

3 AVIDa-SARS-CoV-2: Antigen-VHH Interaction Dataset Produced from Alpaca Immunized with SARS-CoV-2 Spike Proteins

AVIDa-SARS-CoV-2 is an antigen-VHH interaction dataset with 77,003 data samples, comprising 22,002 binding pairs and 55,001 non-binding pairs. The dataset was released under a CC BY-NC 4.0 license and is available at `https://avida-sars-cov-2.cognanous.com`.

3.1 Dataset Generation

AVIDa-SARS-CoV-2 was generated using a method established in our previous study [66]. This section introduces the overall workflow and key concepts underlying our data generation, as shown in Figure 1. Appendix A.2 provides the detailed step-by-step procedures for dataset generation.

Immunization We used the immune system of live alpacas to obtain diverse VHHs that bind to SARS-CoV-2. First, we immunized two alpacas (hereafter referred to as Alpaca P and Alpaca C) that were maternal half-siblings with the 13 types of antigens listed in Table 2. The spike protein of SARS-CoV-2, which protrudes from the virus surface, is a crucial structural component that

https://doi.org/10.52202/079017-3688

Table 2: Summary of antigen types. Appendix A.2 gives more details on each antigen.

Antigen Type	Panning	Description
WT	cell	Wild-type (**WT**) SARS-CoV-2 identified in Wuhan
D614G	cell	Mutant with **D614G** mutation
Alpha	cell, bead	Mutant with representative mutations of **Alpha** variant
Alpha+K417N	cell	Mutant of antigen type “Alpha” with **K417N** mutation
Alpha+E484K	cell	Mutant of antigen type “Alpha” with **E484K** mutation
Beta	cell, bead	Mutant with representative mutations of **Beta** variant
Delta	cell, bead	Mutant with representative mutations of **Delta** variant
Kappa	bead	Mutant with representative mutations of **Kappa** variant
Lambda	bead	Mutant with representative mutations of **Lambda** variant
Omicron	cell, bead	Mutant with representative mutations of **Omicron** (BA.1) variant
PMS	bead	Polymutant spike (**PMS**) protein [58]
S2-domain	bead	**S2-domain** of the WT
OC43	bead	Human coronavirus **OC43** (HCoV-OC43)

facilitates its entry into host cells by binding to receptors on cells. Owing to its crucial role in the infection process, the spike protein is the primary target for antibodies. As the virus evolves over time, mutations in the spike protein that escape the immune response are enriched, and the effectiveness of antibodies to neutralize the virus is reduced. To investigate the effects of mutations in the spike protein, we generated mutants by selecting representative mutations that are effective for immune escape among the mutations observed to date.

Affinity Selection After immunization, an alpaca’s body harbors a small amount of SARS-CoV-2-specific VHHs produced by the immune response and a large amount of VHHs unrelated to SARS-CoV-2. To distinguish between them, we performed affinity selection by biopanning using the spike proteins listed in Table 2 as target molecules. We performed either bead panning, cell panning, or both for each target molecule. For bead panning, the ectodomain of the spike protein was produced by cells, purified, and then combined with beads as bait for panning. For cell panning, the bait was a whole cell overexpressing the full-length spike proteins on the cell membrane with the ectodomain protruding out. Through this process, target-specific VHHs become enriched, while non-specific VHHs are gradually diluted out, ultimately yielding a concentrated sample of target-specific VHHs.

Data Labeling We counted the number of occurrences of each unique VHH sequence in the samples before and after affinity selection by NGS, which reflected the proportion of each VHH in the samples. We then compared the proportions of each VHH before and after affinity selection and labeled VHHs whose proportions significantly increased as “binder” and VHHs whose proportions significantly decreased as “non-binder” on the basis of statistical tests. In addition, VHHs whose proportions did not change significantly, corresponding to about 97% of the total, were excluded from the dataset to improve label reliability. We previously verified the reliability of this labeling method by confirming the binding ability of 20 labeled VHHs in AVIDa-hIL6 [66] using immunofluorescence staining analysis and biolayer interferometry analysis. Furthermore, label reliability was supported by experimental evidence that nine VHHs extracted from AVIDa-SARS-CoV-2 bound to SARS-CoV-2 spike variants, including Omicron [38].

3.2 Dataset Analysis

Binding Sensitivity to Sequence Variation AVIDa-SARS-CoV-2 contains information regarding whether the same VHH sequence binds to each antigen type. The number of unique VHH sequences in AVIDa-SARS-CoV-2 is 36,100 including 14,078 sequences that bind to at least one antigen type. Notably, 427 VHH sequences were labeled as “binder” to specific antigen types but “non-binder” to others. In the case of infectious diseases and malignancies, the target antigen can mutate to escape the immune system or develop tolerance to treatment. If the binding site of an antigen, called an epitope, is mutated, the corresponding antibody will no longer bind to it. Conversely, if an antibody that loses binding activity due to an antigen mutation is identified, the location of the mutation can be assumed to be close to the epitope. Therefore, we further examined the binding activity of these 427 VHH sequences.

First, we clustered 427 VHH sequences using MMseqs2 [63] with 90% sequence identity, resulting in 38 clusters of size two or more. We then extracted 54 VHH sequences from the three clusters in

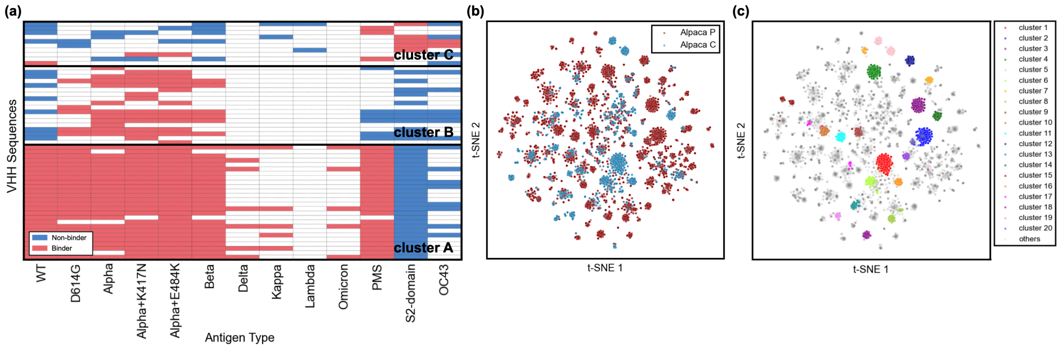

Figure 2: (a) Label visualization for each pair between 54 VHHs in three clusters and antigens. Each cell represents unique VHH-antigen pair. White cells are unlabeled pairs that cannot be identified as "binder" or "non-binder" and are not included in AVIDa-SARS-CoV-2. (b)(c) Two-dimensional representation of binder sequences colored by individuals and clusters. Appendix A.3 provides enlarged versions of (b) and (c).

descending order of cluster size and visualized whether each VHH sequence bound to each antigen type, as shown in Figure 2(a). Focusing on the vertical direction in cluster C, some sequences with over 90% sequence identity can exhibit varying binding abilities against the same antigen. Focusing on the horizontal direction, it is clear that the same VHH has different binding abilities for different antigen types. For example, in cluster B, antigen types from WT to Beta, which differ by only a few amino acids, can alter VHH binding, suggesting that these mutations enhance or inhibit binding. Interestingly, all VHHs in cluster A bind to WT but not to S2-domain, indicating that these VHHs bind to the S1 region of the spike protein. We can also recognize that most of these VHHs cannot be identified as binders for variants such as Delta, Kappa, Lambda, and Omicron, which exactly reflects the immune escape phenomenon in the real world. Therefore, AVIDa-SARS-CoV-2 contains sensitive information in which small amino acid sequence variations of an antibody and antigen can change between binding or non-binding, which should be strongly associated with their binding sites.

Individual Differences in Antigen-specific Antibody Production We compared the differences in SARS-CoV-2-specific VHHs produced by the immune systems of the two alpacas. The number of unique VHH binders for Alpaca P and Alpaca C were 10,487 and 3,651, respectively, of which 60 VHHs were observed in both individuals. We encoded VHH sequences using Kidera factors [23], which represent the physicochemical properties of amino acids in a 10-dimensional vector, and then converted them into two-dimensional (2D) vectors using t-SNE [37]. Figure 2(b) shows a 2D representation of the VHH binders. The data points derived from each individual partially overlapped but predominantly aggregated in distinct regions. Figure 2(c) shows a 2D representation of the VHH binders clustered using MMseqs2 with 95% sequence identity and colored into 20 clusters in descending order of cluster size. This result indicates that the aggregations in 2D space reflect the VHH clusters formed on the basis of sequence identity. For example, cluster 1 (colored red) is composed of VHHs produced from Alpaca C, whereas cluster 2 (colored blue) is composed of VHHs produced from Alpaca P. These results demonstrate that using multiple individuals in dataset generation contributes to enhancing the diversity of antigen-specific VHH sequences.

Differences with AVIDa-hIL6 Building on the findings from the above analysis, we elucidate the differences between AVIDa-SARS-CoV-2 and the previously released AVIDa-hIL6 [66], beyond the target antigens used for immunization. AVIDa-hIL6 used human interleukin-6 (IL-6) mutants produced by artificial point mutations, whereas AVIDa-SARS-CoV-2 used SARS-CoV-2 spike proteins with natural mutations that are more important for antigen-antibody interactions. This allowed AVIDa-SARS-CoV-2 to contain labels that reflect the immune escape phenomenon in the real world, as shown in Figure 2(a). Moreover, AVIDa-hIL6 collected VHHs from one alpaca, whereas AVIDa-SARS-CoV-2 collected them from two alpacas. This increased the diversity of antigen-specific antibodies and provided valuable insights into the sequence differences of antigen-specific antibodies between individuals, as shown in Figure 2(b).

4 VHHCorpus-2M: VHH Sequence Corpus Produced from Alpaca

VHHCorpus-2M is a corpus containing 2,040,988 unique VHH sequences. The corpus was released under a CC BY-NC 4.0 license and is available at `https://vhh-corpus.cognanous.com`.

4.1 Dataset Collection

VHHCorpus-2M is a collection of unique VHH sequences from several datasets generated by the process described in Section 3.1 using target antigens other than the SARS-CoV-2 spike protein, such as the human immunodeficiency virus type 1 (HIV-1) envelope protein, human IL-6, HER2, human histone, transmembrane glycoprotein mucin 1 (MUC1), and gram-negative bacteria. We collected VHH sequences from datasets produced by five alpacas, different from those used in the generation of AVIDa-SARS-CoV-2, to avoid potential data leakage and increase the diversity of VHH sequences. Note that the source datasets include publicly available AVIDa-hIL6 [66] in addition to multiple datasets that have not been published as labeled binding datasets. Importantly, we used only VHH sequences that were identified multiple times by NGS in our corpus to avoid sequencing errors and increase sequence reliability.

4.2 Dataset Analysis

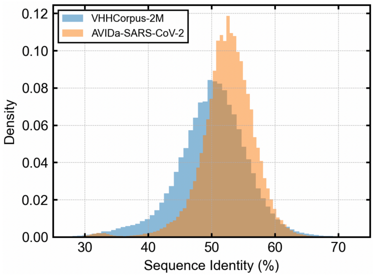

Figure 3: Distribution of pairwise identities of VHH sequences.

The size and diversity of pre-training datasets play an important role in improving the performance of a language model in downstream tasks [51, 33]. VHHCorpus-2M comprises 2,040,988 unique sequences, which is more than 50 times the number of unique sequences in AVIDa-SARS-CoV-2. To examine the degree of sequence diversity in the datasets, we calculated the pairwise sequence identities within each dataset. First, to mitigate the computational complexity, we used MMseqs2 to cluster the VHH sequences with 70% sequence identity within each dataset, resulting in 8,270 and 777 clusters for VHHCorpus-2M and AVIDa-SARS-CoV-2, respectively. We then extracted representative sequences from all clusters and calculated all pairwise sequence identities among these representatives for each dataset. Figure 3 presents the distribution of pairwise sequence identities for each dataset. The distribution of VHHCorpus-2M has a broader peak in regions with a lower sequence identity compared with AVIDa-SARS-CoV-2, indicating a higher sequence diversity. This could be attributed to the fact that VHHCorpus-2M originated from five alpacas, which is more than AVIDa-SARS-CoV-2. Additionally, VHHCorpus-2M includes unlabeled VHH sequences that cannot be labeled as "binder" or "non-binder" for specific antigens, further contributing to its diversity.

5 Benchmarks

5.1 Benchmark Task

Table 3: Numbers of samples in training and test sets.

Dataset	#Samples		
	Binder	Non-binder	Total
Training	15,400	34,285	49,685
Test	6,602	20,716	27,318

To evaluate the performance of various pre-trained language models for antibody discovery, we defined a binary classification task to predict the binding or non-binding of unknown antibodies to 13 antigens using AVIDa-SARS-CoV-2. By leveraging the binding information of diverse VHHs produced from the two alpacas, we used data samples obtained from Alpaca P as the training set and data samples obtained from Alpaca C as the test set. Table 3 lists the number of samples in each set. As shown in Figure 2(b) and (c), the VHH binders derived from different alpacas formed distinct clusters. Therefore, this experimental scenario assumes that we want to explore additional effective antibodies beyond those already observed to bind to a known antigen. This scenario holds significant importance in the development of therapeutic antibodies, given that antibodies with different sequences can bind to different binding sites of antigens, called epitopes. Depending on their binding sites, antibodies may have specific biologically important functions, such as neutralization, inhibition, or activation, and can be extremely useful in drugs.

5.2 Experimental Settings

Baseline Models To fully evaluate the representation capabilities of the language models pre-trained on various training sequence data, we selected the following nine baseline models. (1) **ProtBert** [15] is a BERT-based [13] model pre-trained on 216 million protein sequences in UniRef [65]. (2)(3) **ESM-2** [32] is a model pre-trained on 65 million unique protein sequences in UniRef [65]. We used ESM-2 with 150 and 650 million parameters (hereafter referred to as ESM-2 150M and ESM-2 650M). We adopted ProtBert and ESM-2 to confirm whether pre-training with antibodies, a subset of proteins, is effective for predicting VHH binding. (4) **AbLang-H** [47] is a RoBERTa-based [33] model pre-trained on 14 million heavy chains of antibodies in the OAS database. (5) **AntiBERTa2** [5] is a RoFormer-based [64] model pre-trained using 824 million antibody sequences including paired antibody sequences in the OAS and proprietary database. (6) **AntiBERTa2-CSSP** [5] is a multimodal version of AntiBERTa2 that is further trained on human antibody structures using contrastive sequence-structure pre-training (CSSP). (7) **IgBert** [22] is a model initialized with weights of ProtBert and trained using more than two billion unpaired sequences of light and heavy chains and two million paired sequences in the OAS database. (8) **VHHBERT** is a RoBERTa-based model pre-trained on two million VHH sequences in VHHCorpus-2M. We used the same model parameters as RoBERTa$_{\text{BASE}}$, except that it used positional embeddings with a length of 185 to cover the maximum sequence length of 179 in VHHCorpus-2M. (9) **VHHBERT w/o PT** is a VHHBERT initialized with random weights without pre-training. We adopted this model to confirm the effectiveness of pre-training.

Pre-training As a pre-training corpus for VHHBERT, VHHCorpus-2M was randomly divided into 2,000,000 training sets and 40,988 validation sets. The VHH sequences were tokenized by mapping each of the 20 amino acids to a different token ID and adding special tokens at the beginning and end of the sequence. We used masked language modeling as the pre-training objective. During pre-training, 15% of the residues from each VHH sequence were randomly selected, and of these, 80% were masked, 10% were randomly changed to another residue, and 10% remained unchanged. VHHBERT was pre-trained for 312,500 steps, which equates to 20 epochs, with a batch size of 128 on one NVIDIA Tesla V100 GPU. The resulting VHHBERT is available on the Hugging Face Hub[1].

Fine-tuning As a fine-tuning dataset, we used AVIDa-SARS-CoV-2, which was divided by individual, as shown in Table 3. Figure 4 shows an overview of the experimental setup. AVIDa-SARS-CoV-2 has the amino acid sequences of VHHs and antigens as input features for binding prediction. To obtain each sequence representation, we used the nine aforementioned baseline models for VHHs and the pre-trained protein language model ESM-2 for antigens and extracted the mean of the representations for each amino acid from the last layer in each language model. The sequence representations of the VHHs and antigens were concatenated and utilized as input to a multi-layer perceptron, which was added on top of the two language models as a classification head. Note that we fixed the weights of ESM-2 used for antigens and fine-tuned the classification head and the language model used for VHHs to assess the representation capabilities of antibody language models. We trained the models for 30 epochs with a batch size of 32 on one NVIDIA Tesla V100 GPU. We conducted five repetitive experiments with different random seeds and report the average results and standard derivation.

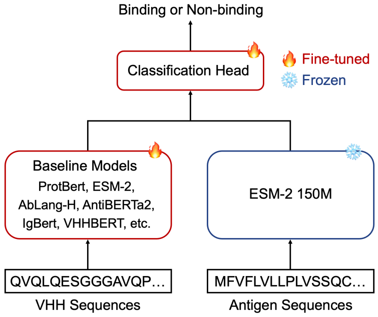

Figure 4: Overview of the experimental setup.

5.3 Results

Table 4 shows the performance comparisons of the baseline models for the VHH-antigen binding prediction. We used precision, recall, F1-score, and area under the precision-recall curve (AUPRC) in addition to accuracy as evaluation metrics because the prediction of antibody binders, which are fewer in number than non-binders, is much more important for drug discovery. VHHBERT w/o PT showed high precision but significantly lower recall than the other models, resulting in the lowest F1-score.

[1] COGNANO/VHHBERT: `https://huggingface.co/COGNANO/VHHBERT`

Table 4: Performance comparisons of baseline models for VHH-antigen binding prediction. Best performance is highlighted in bold.

Model	Accuracy	Precision	Recall	F1-score	AUPRC
ProtBert	0.803 ± 0.012	0.602 ± 0.036	0.564 ± 0.046	0.580 ± 0.023	0.532 ± 0.073
ESM-2 150M	0.801 ± 0.010	0.607 ± 0.034	0.514 ± 0.036	0.555 ± 0.021	0.531 ± 0.047
ESM-2 650M	0.822 ± 0.020	0.682 ± 0.083	0.540 ± 0.048	0.598 ± 0.023	0.584 ± 0.069
AbLang-H	0.828 ± 0.004	0.753 ± 0.033	0.430 ± 0.017	0.547 ± 0.005	0.589 ± 0.018
AntiBERTa2	0.851 ± 0.007	0.769 ± 0.044	0.551 ± 0.021	0.641 ± 0.008	0.660 ± 0.018
AntiBERTa2-CSSP	**0.854** ± **0.007**	0.773 ± 0.030	0.565 ± 0.014	**0.652** ± **0.014**	**0.690** ± **0.011**
IgBert	0.845 ± 0.007	0.741 ± 0.045	0.558 ± 0.045	0.634 ± 0.018	0.610 ± 0.044
VHHBERT	0.823 ± 0.011	0.658 ± 0.042	**0.567** ± **0.025**	0.608 ± 0.012	0.650 ± 0.025
VHHBERT w/o PT	0.831 ± 0.003	**0.811** ± **0.024**	0.392 ± 0.010	0.528 ± 0.008	0.624 ± 0.008

This result indicates the effectiveness of pre-training on protein and antibody sequences in predicting VHH binding. AntiBERTa2, AntiBERTa2-CSSP, IgBert, and VHHBERT pre-trained on antibody sequences outperformed ESM-2 and ProtBert pre-trained on protein sequences in accuracy, F1-score, and AUPRC. This is consistent with previous studies [27, 70, 5] that reported that using antibodies for pre-training, rather than general proteins, contributes to the performance of antibody-specific tasks. In general, the antigen to which an antibody binds is determined by the amino acid sequence in CDRs. Because CDRs are highly variable owing to mechanisms such as immunoglobulin gene rearrangement and somatic hypermutation [9], they do not follow the evolutionary information stored in general protein sequences [67]. Because of these differences in evolutionary processes, pre-training with antibody sequences should be effective in predicting VHH binding.

AntiBERTa2, AntiBERTa2-CSSP, and IgBert outperformed the other models in terms of accuracy and F1-score, suggesting that pre-training with a larger number of antibody sequences contributes to the generalization to unknown antibody clusters. Interestingly, additional pre-training of AntiBERTa2-CSSP using human antibody structures contributed to improved performance in predicting VHH-antigen binding. However, the highest F1-score of AntiBERTa2-CSSP remains at approximately 65%; therefore, there is still room for performance improvement for practical drug discovery applications. Although VHHBERT was pre-trained with significantly fewer antibody sequences than the other pre-trained antibody language models, its F1-score was higher than AbLang-H and close to IgBert. This result can probably be attributed to differences in the sequence patterns between conventional antibodies and VHHs. Specifically, the average length of CDR3, which is the most important for antigen recognition, is approximately 1.5 times longer for VHHs than for conventional antibodies [68, 17]. Moreover, the genetic sequences of antibodies differ between species, resulting in differences in amino acid sequences, even in non-variable regions [34, 62]. In conclusion, these insights obtained through benchmarks underscore the significance of AVIDa-SARS-CoV-2 as a useful benchmark for assessing the representation capabilities of antibody language models for binding prediction, thereby promoting the advancement of AI-assisted antibody discovery.

6 Discussion

6.1 VHH-specific Language Models

Recent remarkable progress in language models has led to the active development of domain-specific language models in various application fields [6, 16, 36]. Here, we discuss the significance of building VHH-specific language models from the perspective of drug discovery. VHHs have recently attracted attention as therapeutic agents because of their small size, high stability, good human tolerability, and relative ease of production [20, 19]. Furthermore, VHHs possess favorable properties for the construction of large-scale language models. First, VHHs have a simple structure consisting of only heavy chains, which allows for easier identification of full-length amino acid sequences using DNA sequencing technologies. Second, VHH acts as a single functional unit, meaning that VHH sequences contain all the information necessary for the function of an antibody against an antigen. In contrast, conventional antibodies are composed of two pairs of heavy and light chains, and they function as a single functional unit by combining heavy and light chains. Therefore, paired sequences should ideally be used as the input data for language models. However, it is difficult to construct a large-scale database of paired sequences because obtaining them requires time-consuming experiments. Indeed, the number of paired sequences recorded in the OAS database is approximately one thousand times less than that of unpaired sequences. Accordingly, the existing antibody language models are trained primarily on unpaired heavy- and/or light-chain sequences, ignoring the effects of their counterpart chains. Although AntiBERTa2 and IgBert used paired sequences in their training data, the majority of the training data consists of unpaired sequences. The advantages of VHH over conventional antibodies will facilitate the construction of large-scale databases that are meaningful for therapeutic antibody discovery and pave the way for future construction of practical VHH-specific language models.

6.2 Negative Societal Impacts

The VHH binders in AVIDa-SARS-CoV-2 have the potential to be useful in COVID-19 therapeutics. In addition, using our dataset to develop predictive models for binding to SARS-CoV-2 variants may accelerate the development of therapeutics against emerging variants of concern (VOCs). To reap these benefits, there is a possibility that third-party organizations could use our dataset for commercial purposes. Because this creates the risk of future conflicts of interest between third-party organizations, we have prohibited commercial use by licensing the dataset. Furthermore, antibodies usually act as inhibitors or neutralizers of their target antigen, but some, although relatively rare, can stimulate or enhance the function of the target [57, 29]. Even if antibodies that bind to spike proteins can be identified or predicted, the possibility of promoting infection cannot be excluded. Therefore, validation experiments and clinical trials must be conducted to confirm the usefulness of each VHH.

6.3 Limitations and Future Work

Our dataset potentially contains data biases derived from the specific alpacas used for dataset generation. As shown in Figure 2(b), each individual produced biased SARS-CoV-2-specific VHHs. This limitation may reduce the generalization performance of the models trained on our dataset, potentially hindering their practical application in antibody discovery. The best way to address this limitation is to generate datasets from multiple individuals with different VHH gene sequences at multiple times of immunization. This would mitigate data biases in the dataset and help a model to understand the universal knowledge of antigen-antibody interactions. However, realistically, there are limitations in creating datasets under various conditions, owing to cost and time constraints. Therefore, it is also necessary to develop models that can overcome data biases. Our benchmark task serves as a valuable benchmark for evaluating whether models can overcome individual-induced biases, and to our knowledge, no other such datasets exist. In the future, in addition to generating and publishing more diverse datasets, we plan to research further model architectures and pre-training methods to achieve a high generalization performance in real-world antibody discovery tasks.

7 Conclusion

In this study, we introduced AVIDa-SARS-CoV-2, a labeled dataset of SARS-CoV-2-VHH interactions, and VHHCorpus-2M, which contains over two million VHH sequences, providing novel datasets for the evaluation and pre-training of antibody language models. In addition, we developed a VHH-specific language model, VHHBERT, pre-trained on VHHCorpus-2M and reported benchmark results for binding prediction using existing general protein and antibody-specific language models. We envision that the availability of AVIDa-SARS-CoV-2 and VHHCorpus-2M will facilitate further research on antibody language models and their application in therapeutic antibody discovery.

https://doi.org/10.52202/079017-3688

Acknowledgments and Disclosure of Funding

This study was supported by funding from GREEN CORE, LTD. (Tokyo, Japan), Yamato scientific group holdings Inc. (Tokyo, Japan), and POPURI Pharmacy Co., Ltd. (Kyoto, Japan) and subsidies from KYOTO Industrial Support Organization 21 (Sangakukou no Mori). This study was also supported by the Ministry of Economy, Trade and Industry (METI) R&D Support Program for Growth-oriented Technology SMEs Grant Number JPJ005698.

We thank the Japan External Trade Organization (JETRO) for supporting COGNANO Inc. We also thank Tomohisa Oda from COGNANO Inc. and SAKURA internet Inc. for developing the websites for the publication of AVIDa-SARS-CoV-2 and VHHCorpus-2M.

References

[1] Al Ojaimi, Y., Blin, T., Lamamy, J., Gracia, M., Pitiot, A., Denevault-Sabourin, C., Joubert, N., Pouget, J.P., Gouilleux-Gruart, V., Heuzé-Vourc'h, N., et al.: Therapeutic antibodies–natural and pathological barriers and strategies to overcome them. Pharmacology & Therapeutics **233**, 108022 (2022)

[2] Alejandra, W.P., Irene, J.P.M., Antonio, G.S.F., Patricia, R.G.R., Elizabeth, T.A., Pablo, A.A.J., Rebeca, G.V.: Production of monoclonal antibodies for therapeutic purposes: A review. International Immunopharmacology **120**, 110376 (2023)

[3] Aronesty, E.: Comparison of sequencing utility programs. The open bioinformatics journal **7**(1) (2013)

[4] Bai, G., Sun, C., Guo, Z., Wang, Y., Zeng, X., Su, Y., Zhao, Q., Ma, B.: Accelerating antibody discovery and design with artificial intelligence: recent advances and prospects. Seminars in Cancer Biology **95**, 13–24 (2023)

[5] Barton, J., Galson, J.D., Leem, J.: Enhancing antibody language models with structural information. In: Machine Learning for Structural Biology Workshop, NeurIPS (2023)

[6] Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)

[7] Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014)

[8] Burbach, S.M., Briney, B.: Improving antibody language models with native pairing. Patterns **5**(5) (2024)

[9] Chi, X., Li, Y., Qiu, X.: V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology **160**(3), 233–247 (2020)

[10] Chungyoun, M., Ruffolo, J.A., Gray, J.J.: FLAb: Benchmarking deep learning methods for antibody fitness prediction. In: Machine Learning for Structural Biology Workshop, NeurIPS (2023)

[11] Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al.: Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009)

[12] Deszyński, P., Młokosiewicz, J., Volanakis, A., Jaszczyszyn, I., Castellana, N., Bonissone, S., Ganesan, R., Krawczyk, K.: INDI—integrated nanobody database for immunoinformatics. Nucleic Acids Research **50**(D1), D1273–D1281 (2022)

[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[14] Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., Deane, C.M.: SAbDab: the structural antibody database. Nucleic acids research **42**(D1), D1140–D1146 (2014)

[15] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al.: ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence **44**(10), 7112–7127 (2022)

https://doi.org/10.52202/079017-3688

[16] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1), 1–23 (2021)

[17] Henry, K.A., MacKenzie, C.R.: Antigen recognition by single-domain antibodies: structural latitudes and constraints. mAbs **10**(6), 815–826 (2018)

[18] Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., Kim, P.S.: Efficient evolution of human antibodies from general protein language models. Nature Biotechnology **42**, 275–283 (2024)

[19] Jin, B.k., Odongo, S., Radwanska, M., Magez, S.: Nanobodies: A review of generation, diagnostics and therapeutics. International Journal of Molecular Sciences **24**(6), 5994 (2023)

[20] Jovčevska, I., Muyldermans, S.: The therapeutic potential of nanobodies. BioDrugs **34**(1), 11–26 (2020)

[21] Kandari, D., Bhatnagar, R.: Antibody engineering and its therapeutic applications. International Reviews of Immunology **42**(2), 156–183 (2023)

[22] Kenlay, H., Dreyer, F.A., Kovaltsuk, A., Miketa, D., Pires, D., Deane, C.M.: Large scale paired antibody language models. arXiv preprint arXiv:2403.17889 (2024)

[23] Kidera, A., Konishi, Y., Oka, M., Ooi, T., Scheraga, H.A.: Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry **4**, 23–55 (1985)

[24] Kim, J., McFee, M., Fang, Q., Abdin, O., Kim, P.M.: Computational and artificial intelligence-based methods for antibody development. Trends in Pharmacological Sciences **44**(3), 175–189 (2023)

[25] Koenig, P., Lee, C.V., Walters, B.T., Janakiraman, V., Stinson, J., Patapoff, T.W., Fuh, G.: Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. Proceedings of the National Academy of Sciences **114**(4), E486–E495 (2017)

[26] Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., Krawczyk, K.: Observed Antibody Space: a resource for data mining next-generation sequencing of antibody repertoires. The Journal of Immunology **201**(8), 2502–2509 (2018)

[27] Leem, J., Mitchell, L.S., Farmery, J.H., Barton, J., Galson, J.D.: Deciphering the language of antibodies using self-supervised learning. Patterns **3**(7), 100513 (2022)

[28] Leinonen, R., Sugawara, H., Shumway, M.: The Sequence Read Archive. Nucleic Acids Research **39**(suppl_1), D19–D21 (2011)

[29] Leitner, J., Egerer, R., Waidhofer-Söllner, P., Grabmeier-Pfistershammer, K., Steinberger, P.: FcγR requirements and costimulatory capacity of Urelumab, Utomilumab, and Varlilumab. Frontiers in Immunology **14** (2023)

[30] Li, G., Hilgenfeld, R., Whitley, R., De Clercq, E.: Therapeutic strategies for COVID-19: progress and lessons learned. Nature Reviews Drug Discovery **22**, 449–475 (2023)

[31] Li, X., Duan, X., Yang, K., Zhang, W., Zhang, C., Fu, L., Ren, Z., Wang, C., Wu, J., Lu, R., et al.: Comparative analysis of immune repertoires between Bactrian camel's conventional and heavy-chain antibodies. PLoS One **11**(9), e0161801 (2016)

[32] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[33] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[34] de Los Rios, M., Criscitiello, M.F., Smider, V.V.: Structural and genetic diversity in antibody repertoires from diverse species. Current Opinion in Structural Biology **33**, 27–41 (2015)

[35] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

https://doi.org/10.52202/079017-3688

[36] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics **23**(6), bbac409 (2022)

[37] Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[38] Maeda, R., Fujita, J., Konishi, Y., Kazuma, Y., Yamazaki, H., Anzai, I., Watanabe, T., Yamaguchi, K., Kasai, K., Nagata, K., et al.: A panel of nanobodies recognizing conserved hidden clefts of all SARS-CoV-2 spike variants including Omicron. Communications Biology **5**, 669 (2022)

[39] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17**(1), 10–12 (2011)

[40] Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S.M., Ehling, R.A., Bonati, L., Dahinden, J., Gainza, P., et al.: Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nature Biomedical Engineering **5**, 600–612 (2021)

[41] Melnyk, I., Chenthamarakshan, V., Chen, P.Y., Das, P., Dhurandhar, A., Padhi, I., Das, D.: Reprogramming pretrained language models for antibody sequence infilling. In: International Conference on Machine Learning. pp. 24398–24419. PMLR (2023)

[42] Miroshnikov, K.A., Marusich, E.I., Cerritelli, M.E., Cheng, N., Hyde, C.C., Steven, A.C., Mesyanzhinov, V.V.: Engineering trimeric fibrous proteins based on bacteriophage T4 adhesins. Protein Engineering, Design and Selection **11**(4), 329–332 (1998)

[43] Mroczek, E.S., Ippolito, G.C., Rogosch, T., Hoi, K.H., Hwangpo, T.A., Brand, M.G., Zhuang, Y., Liu, C.R., Schneider, D.A., Zemlin, M., et al.: Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Frontiers in immunology **5** (2014)

[44] Nakanishi, T., Tsumoto, K., Yokota, A., Kondo, H., Kumagai, I.: Critical contribution of VH–VL interaction to reshaping of an antibody: The case of humanization of anti-lysozyme antibody, HyHEL-10. Protein Science **17**(2), 261–270 (2008)

[45] Nijkamp, E., Ruffolo, J.A., Weinstein, E.N., Naik, N., Madani, A.: ProGen2: exploring the boundaries of protein language models. Cell Systems **14**(11), 968–978 (2023)

[46] Olsen, T.H., Boyles, F., Deane, C.M.: Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science **31**(1), 141–146 (2022)

[47] Olsen, T.H., Moal, I.H., Deane, C.M.: AbLang: an antibody language model for completing antibody sequences. Bioinformatics Advances **2**(1), vbac046 (2022)

[48] Pallesen, J., Wang, N., Corbett, K.S., Wrapp, D., Kirchdoerfer, R.N., Turner, H.L., Cottrell, C.A., Becker, M.M., Wang, L., Shi, W., et al.: Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proceedings of the National Academy of Sciences **114**(35), E7348–E7357 (2017)

[49] Porebski, B.T., Balmforth, M., Browne, G., Riley, A., Jamali, K., Fürst, M.J., Velic, M., Buchanan, A., Minter, R., Vaughan, T., et al.: Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. Nature biomedical engineering **8**, 214–232 (2024)

[50] Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., Bitton, D.A.: BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. mAbs **14**(1), 2020203 (2022)

[51] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog (2019)

[52] Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: MSA transformer. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8844–8856. PMLR (2021)

[53] Raybould, M.I., Kovaltsuk, A., Marks, C., Deane, C.M.: CoV-AbDab: the coronavirus antibody database. Bioinformatics **37**(5), 734–735 (2021)

https://doi.org/10.52202/079017-3688

[54] Rice, P., Longden, I., Bleasby, A.: EMBOSS: the european molecular biology open software suite. Trends in genetics **16**(6), 276–277 (2000)

[55] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences **118**(15), e2016239118 (2021)

[56] Ruffolo, J.A., Gray, J.J., Sulam, J.: Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv preprint arXiv:2112.07782 (2021)

[57] Schardt, J.S., Jhajj, H.S., O'Meara, R.L., Lwo, T.S., Smith, M.D., Tessier, P.M.: Agonist antibody discovery: Experimental, computational, and rational engineering approaches. Drug Discovery Today **27**(1), 31–48 (2022)

[58] Schmidt, F., Weisblum, Y., Rutkowska, M., Poston, D., DaSilva, J., Zhang, F., Bednarski, E., Cho, A., Schaefer-Babajew, D.J., Gaebler, C., et al.: High genetic barrier to SARS-CoV-2 polyclonal neutralizing antibody escape. Nature **600**, 512–516 (2021)

[59] Shanehsazzadeh, A., Bachas, S., McPartlon, M., Kasun, G., Sutton, J.M., Steiger, A.K., Shuai, R., Kohnert, C., Rakocevic, G., Gutierrez, J.M., et al.: Unlocking de novo antibody design with generative artificial intelligence. bioRxiv (2023)

[60] Shen, W., Le, S., Li, Y., Hu, F.: SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS one **11**(10), e0163962 (2016)

[61] Shuai, R.W., Ruffolo, J.A., Gray, J.J.: Generative language modeling for antibody design. In: Machine Learning for Structural Biology Workshop, NeurIPS (2021)

[62] Sinkora, M., Stepanova, K., Butler, J.E., Sinkora Jr, M., Sinkora, S., Sinkorova, J.: Comparative aspects of immunoglobulin gene rearrangement arrays in different species. Frontiers in Immunology **13** (2022)

[63] Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**, 1026–1028 (2017)

[64] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[65] Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., Consortium, U.: UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31**(6), 926–932 (2015)

[66] Tsuruta, H., Yamazaki, H., Maeda, R., Tamura, R., Wei, J.N., Mariet, Z., Phloyphisut, P., Shimokawa, H., Ledsam, J.R., Colwell, L., Imura, A.: AVIDa-hIL6: A large-scale VHH dataset produced from an immunized alpaca for predicting antigen-antibody interactions. In: Advances in Neural Information Processing Systems 36 (2023)

[67] Vishwakarma, P., Vattekatte, A.M., Shinada, N., Diharce, J., Martins, C., Cadet, F., Gardebien, F., Etchebest, C., Nadaradjane, A.A., de Brevern, A.G.: VHH structural modelling approaches: A critical review. International Journal of Molecular Sciences **23**(7), 3721 (2022)

[68] Vu, K.B., Ghahroudi, M.A., Wyns, L., Muyldermans, S.: Comparison of llama VH sequences from conventional and heavy chain antibodies. Molecular immunology **34**(16-17), 1121–1131 (1997)

[69] Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D.: Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell **181**(2), 281–292 (2020)

[70] Wang, D., Fei, Y., Zhou, H.: On pre-training language model for antibody. In: The Eleventh International Conference on Learning Representations (2023)

[71] Warszawski, S., Borenstein Katz, A., Lipsh, R., Khmelnitsky, L., Ben Nissan, G., Javitt, G., Dym, O., Unger, T., Knop, O., Albeck, S., et al.: Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. PLOS Computational Biology **15**(8), e1007207 (2019)

[72] Wilman, W., Wróbel, S., Bielska, W., Deszynski, P., Dudzic, P., Jaszczyszyn, I., Kaniewski, J., Młokosiewicz, J., Rouyan, A., Satława, T., et al.: Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. Briefings in Bioinformatics **23**(4), bbac267 (2022)

[73] Zhang, L., Leng, Q., Mixson, A.J.: Alteration in the IL-2 signal peptide affects secretion of proteins in vitro and in vivo. The Journal of Gene Medicine **7**(3), 354–365 (2005)

[74] Zhou, G., Zhao, Q.: Perspectives on therapeutic neutralizing antibodies against the novel coronavirus SARS-CoV-2. International Journal of Biological Sciences **16**(10), 1718–1723 (2020)

[75] Zhou, T., Lynch, R.M., Chen, L., Acharya, P., Wu, X., Doria-Rose, N.A., Joyce, M.G., Lingwood, D., Soto, C., Bailer, R.T., et al.: Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. Cell **161**(6), 1280–1292 (2015)

[76] Zhou, T., Zhu, J., Wu, X., Moquin, S., Zhang, B., Acharya, P., Georgiev, I.S., Altae-Tran, H.R., Chuang, G.Y., Joyce, M.G., et al.: Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. Immunity **39**, 245–258 (2013)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.3.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.2.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3, Section 4, and Appendix A.4 for links to data and code. The code contains a readme with instructions to reproduce the experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1, Section 5.2, and Appendix A.4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.2 and Appendix A.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix A.4.
   (b) Did you mention the license of the assets? [Yes] See Appendix A.4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 3 and Section 4.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/079017-3688

A Appendix

A.1 Ethics Statement for Animal Experiments

All animal experiments on the alpacas were conducted in accordance with the KYODOKEN Institute for Animal Science Research and Development (Kyoto, Japan) and the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines[2]. Veterinarians performed breeding, health maintenance, and immunization by adhering to the published Guidelines for Proper Conduct of Animal Experiments by the Science Council of Japan. The KYODOKEN Institutional Animal Care and Use Committee approved the protocols for these studies (approval number 20200312).

Our data generation method uses animal models immunized with a target protein that could potentially harm the animals, such as a particular toxin, pathogen, or allergen. Hence, the risk to the animal should be minimized by treating the immune source to inactivate or detoxify it.

A.2 Dataset Generation

Here, we describe the detailed experimental procedures and conditions.

Immunization We immunized two alpacas, Alpaca P and Alpaca C, with various purified recombinant SARS-CoV-2 spike trimers: WT-cryo, the ectodomain of the wild-type (WT) SARS-CoV-2 spike protein analyzed by cryo-EM [69], whose S1/S2 region is altered to avoid cleavage by furin; WT-ecto, the restored version of the furin site of the WT-cryo; WT, the full-length of the SARS-CoV-2 spike protein (GenBank: QHD43416); and S2-domain, the S2 domain of the WT. The antigen cocktail mixture, emulsified in complete Freund's adjuvant, was injected subcutaneously into the two alpacas nine times at two-week intervals. Lymph node and blood samples were collected multiple times, resulting in a total of 27 libraries. After this period, we continued immunization with various SARS-CoV-2 spike proteins harboring representative mutations for the Alpha, Beta, Gamma, Delta, Kappa, Lambda, and Omicron variants. As a result, we performed eight additional immunizations with these variants and generated four additional libraries from the final harvest samples.

Phage Library Construction Peripheral blood mononuclear cells (PBMCs) were obtained from blood samples by sucrose density gradient centrifugation using Ficoll (Nacalai Tesque, Kyoto, Japan). The lymph nodes and PBMC samples were washed with phosphate-buffered saline (PBS, Nacalai Tesque) and suspended in an RNAlater solution (Thermo Fisher Scientific K.K., Tokyo, Japan). Total RNA was isolated from these samples by using Direct-Zol RNA MiniPrep (Zymo Research, Irvine, CA). Complementary DNA was synthesized from 1 μg of total RNA as a template by using random hexamer primers and SuperScript II reverse transcriptase (Thermo Fisher Scientific K.K.). The coding regions of the VHH domain were amplified using LA Taq polymerase (TAKARA Bio Inc., Shiga, Japan) with two PAGE-purified primers (CALL001, 5'-GTCCTGGCTGCTCTTCTACAAGG-3' and CALL002, 5'-GGTACGTGCTGTTGAACTGTTCC-3'), and they were separated on a 1.5% low-melting-temperature agarose gel (Lonza Group AG, Basel, Switzerland). Approximately 700 base-pair bands were extracted using a QIAquick Gel Extraction Kit (Qiagen K.K., Tokyo, Japan). Nested PCR was performed to amplify the VHH genes by using two primers that contained flanking PstI (forward) and BstEII (reverse) restriction sites to enable cloning into the pMES4 phagemid vector with a C-terminal His-tag. Electroporation-competent Escherichia coli TG1 cells (Agilent Technologies Japan, Ltd., Tokyo, Japan) were transformed with the ligated plasmids under chilled conditions (Bio-Rad Laboratories, Inc., Hercules, CA). The library densities were monitored and maintained at >10^7 colony-forming units per microliter with limiting dilution. Colonies from 8 mL of cultured cells were harvested, pooled, and reserved in frozen glycerol stock as a mother library. Thus, the 31 phagemid libraries were designated as the mother libraries.

Affinity Selection We adopted two methods for affinity selection: the use of purified ectodomains coated on beads (the bead panning method) and the use of cultured cells overexpressing full-length spike proteins (the cell panning method). This is because we think that the bead panning method tends to give clearer signals than the cell panning method, while the latter reflects a more native three-dimensional structure as multimers. As shown in Table 5, the antigens used in this procedure contain various combinations of mutations at clinically important amino acids that belong to the receptor

[2] ARRIVE guidelines: `https://arriveguidelines.org`

https://doi.org/10.52202/079017-3688

Table 5: Details of antigen types used for dataset generation. Amino acid sequence for each antigen type is available at `https://huggingface.co/datasets/COGNANO/AVIDa-SARS-CoV-2`.

Antigen Type	Panning	Mutation	Description
WT	cell	-	Wild-type (WT) SARS-CoV-2 identified in Wuhan with a C9 tag at the C-terminus.
D614G	cell	D614G	Mutant with D614G mutation with a C9 tag at the C-terminus.
Alpha	cell	N501Y, D614G	Mutant with representative mutations of Alpha variant with a C9 tag at the C-terminus.
Alpha	bead	N501Y, D614G, K986P, V987P	Mutant with representative mutations of Alpha variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
Alpha+K417N	cell	K417N, N501Y, D614G	Mutant of antigen type "Alpha" with K417N mutation with a C9 tag at the C-terminus.
Alpha+E484K	cell	E484K, N501Y, D614G	Mutant of antigen type "Alpha" with E484K mutation with a C9 tag at the C-terminus.
Beta	cell	K417N, E484K, N501Y, D614G	Mutant with representative mutations of Beta variant with a C9 tag at the C-terminus.
Beta	bead	K417N, E484K, N501Y, D614G, K986P, V987P	Mutant with representative mutations of Beta variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
Delta	cell	L452R, T478K, D614G	Mutant with representative mutations of Delta variant with a C9 tag at the C-terminus.
Delta	bead	L452R, T478K, D614G	Mutant with representative mutations of Delta variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
Kappa	bead	L452R, E484Q, D614G, K986P, V987P	Mutant with representative mutations of Kappa variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
Lambda	bead	G75V, T76I, S247_D253del, L452Q, F490S, K986P, V987P	Mutant with representative mutations of Lambda variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
Omicron	cell	A67V, H69del, V70del, T95I, G142D, V143_Y145del, N211I, L212V, V213P, R214E, G339D, S371L, S373P, S735F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F	Mutant with representative mutations of Omicron (BA.1) variant with a C9 tag at the C-terminus.
Omicron	bead	A67V, H69del, V70del, T95I, G142D, V143_Y145del, N211I, L212V, V213P, R214E, G339D, S371L, S373P, S735F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, K986P, V987P	Mutant with representative mutations of Omicron (BA.1) variant with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
PMS	bead	L18F, V47M, H69del, V70del, D80A, Y145del, D215G, L242_L244del, W258R, R346S, K417N, N440K, V445E, L455R, A475V, E484K, N501Y, D614G, A701V, P792H, N801D, K986P, V987P	Polymutant spike (PMS) protein [58] with a 6×His tag at the C-terminus. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
S2-domain	bead	K986P, V987P	The S2 domain of the WT with a 6×His tag at the C-terminus. The first thirty amino acids are replaced by the IL-2 secretion signal peptide [73] and a linker sequence. The PP mutation stabilizes the trimer structure [48]. The sequence after the transmembrane domain is replaced by the foldon trimerization motif [42].
OC43	bead	-	Human coronavirus OC43 (HCoV-OC43) with a 6×His tag at the C-terminus.

binding domain (RBD) such as K417, L452, T478, E484 and N501. For the bead panning method, K986P and V987P substitutions as well as the foldon trimerization motif [42] at the C-terminus were introduced into the antigens. To distinguish nonspecific signals, (i) mock, (ii) APOBEC3G, (iii) homemade IgM were used as controls. Target antigens were coupled to N-hydroxysuccinimide (NHS)-activated magnet beads (Dynabeads, Thermo). One round of biopanning was performed using each target protein-coated magnet beads in 50 mM of phosphate buffer (pH 7.4) containing 1% n-dodecyl-β-D-maltopyranoside (DDM: Nacalai), 0.1% 3-[(3-cho-lamidopropyl)dimethylammonio]-1-propane sulfonate (CHAPS: Nacalai), 0.001% cholesterol hydrogen succinate (CHS: Tokyo Chemical Industry Co., Ltd. (TCI), Tokyo, Japan), 0.1% LMNG (Anatrace, Maumee, OH), and 500 mM of NaCl. After three washes with the same buffer, the remaining phages bound to the beads were eluted with a trypsin-ethylenediaminetetraacetic acid (EDTA, Nacalai Tesque) solution at room temperature for 30 minutes. For the cell panning method, one round of biopanning was performed using PFA-fixed cultured cells. Phages dissolved in PBS containing 0.05% Tween 20, 0.5 mM of NaCl, and 0.3% bovine serum albumin (BSA, Nacalai) were incubated with PFA-fixed cells expressing target spike proteins for 30 minutes at room temperature. The cells were washed three times with PBS containing 0.05% Tween 20, 0.5 mM of NaCl, and 0.3% BSA; the remaining phages bound to the cells were eluted with a trypsin-ethylenediaminetetraacetic acid (EDTA) solution (Nacalai) for 30 minutes at room temperature with gentle agitation. For both panning methods, the eluate was neutralized with a PBS-diluted protein inhibitor cocktail (cOmplete, EDTA-free, protease inhibitor cocktail tablets, Roche Diagnostics GmbH, Mannheim, Germany) and used to infect electroporation-competent cells. The infected cells were cultured in LB Miller broth containing 100 μg/mL of ampicillin (Nacalai Tesque) at 37 °C overnight. The genes of the phagemids selected by biopanning were collected with a QIAprep Miniprep Kit (Qiagen), amplified by PCR, and purified using AMPure XP beads (Beckman Coulter, High Wycombe, UK). Then, dual-indexed libraries were prepared and sequenced on an Illumina MiSeq (Illumina, San Diego, CA) by using a MiSeq Reagent Kit v3 with paired-end 300-bp reads (Bioengineering Lab. Co., Ltd., Kanagawa, Japan).

Sequence Analysis Approximately 100,000 paired reads for each library were generated by NGS analysis. The raw read data were trimmed to remove the adaptor sequence by using Cutadapt v3.5 [39] and to remove low-quality reads by using Trimmomatic v0.39 [7]. The remaining paired reads were merged using fastq-join [3], and then the VHH coding sequences were extracted using SeqKit v2.2.0 [60]. The DNA sequences were translated to amino acid sequences with EMBOSS v6.6.0 [54], and the VHH sequences were cropped from start to stop codon. Finally, each phagemid library was converted to a FASTA file containing tens of thousands of VHH sequences.

Data Labeling Data labeling was performed using our proposed method [66], whose reliability was fully verified using immunofluorescence staining analysis and biolayer interferometry analysis. The code for the data labeling is available at `https://github.com/cognano/AVIDa-SARS-CoV-2`. Our labeling method determines whether a VHH binds to each antigen by applying a statistical test for differences in the proportions of each VHH in a library before and after panning. Let p_1 and p_2 denote the population proportions of a specific VHH in the libraries before and after panning. Moreover, let n_1 and n_2 denote the libraries' total read counts before and after panning, respectively, and let x_1 and x_2 denote the read counts of a specific VHH in the libraries. Then, the respective sample proportions of a specific VHH in each library are $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$. Given that the minimum value of all possible n_1 and n_2 was over 5,000, we assumed that $\hat{p}_1$ and $\hat{p}_2$ follow normal distributions with mean p_1 and p_2 and variance $\frac{p_1(1-p_1)}{n_1}$ and $\frac{p_2(1-p_2)}{n_2}$, respectively, according to the central limit theorem. Furthermore, the difference in the proportions $\hat{p}_1 - \hat{p}_2$ can also be approximated by a normal distribution due to the reproductive property of the normal distribution. Thus, the test statistic Z under null hypothesis $H_0 : p_1 = p_2$ was calculated as follows.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{1}$$

where p is the pooled proportion calculated as $p = \frac{x_1+x_2}{n_1+n_2}$. The p-value of Z was calculated using the standard normal distribution. In the same way, p-values were calculated for all VHH-target pairs in the sublibraries with respect to the corresponding mother libraries. We adopted the smallest p-value, indicating the most significant difference in proportion, among identical VHH-target pairs. If a specific VHH's proportion in a sublibrary increased from the proportion in the corresponding mother library and the p-value was 0.05 or less (our chosen significance level), the VHH-target pair was labeled with "binder." Similarly, if the proportion decreased and the p-value was 0.05 or less, the pair was labeled with "non-binder." Finally, if the p-value exceeded 0.05, the pair was labeled with "non-significant." This label was excluded from AVIDa-SARS-CoV-2.

The results of biological experiments always contain background noise, such as that due to binding to contaminating proteins. Therefore, we applied a novel noise reduction algorithm to avoid false positives and improve label reliability. We reconfirmed VHHs labeled as a "binder" to any of the antigen types by comparing the labels to negative control samples under the following conditions.

1. If the VHH was a non-binder to a negative control sample, the label remained "binder."
2. If the VHH was a binder to a negative control sample, the label was reassigned from "binder" to "noise" because of possible false positives. This label was excluded from AVIDa-SARS-CoV-2.
3. If the VHH was "non-significant" with respect to a negative control sample, the ratio of the p-value of the negative control sample to that of the antigen was compared to $10^{2.5}$. This value was empirically determined by an author (a biologist) according to feedback from biological experiments in our previous studies [38].
    (a) If the ratio of p-values was below $10^{2.5}$, the label was reassigned from "binder" to "non-significant" because of possible false positives.
    (b) If the ratio of p-values was $10^{2.5}$ or more, the label remained "binder."

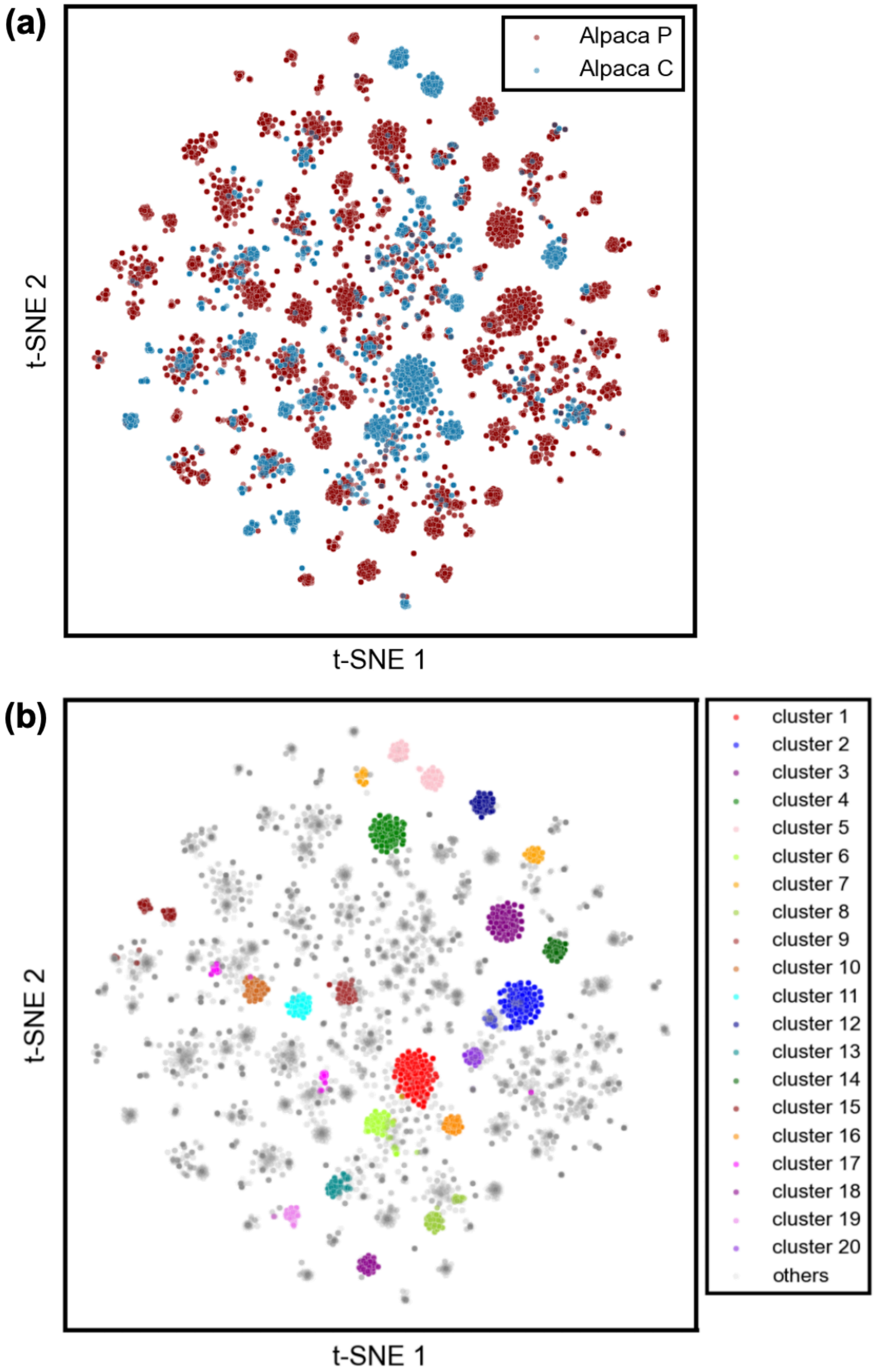

Figure 5: Two-dimensional representation of binder sequences colored by (a) individuals and (b) clusters.

Table 6: Comparison of model parameters of baseline models. “M” stands for million.

Parameters	ProtBert	ESM-2 150M	ESM-2 650M	AbLang-H	AntiBERTa2 AntiBERTa2-CSSP	IgBert	VHHBERT
Number of layers	30	30	33	12	16	30	12
Number of attention heads	16	20	20	12	16	16	12
Embedding dimension	1024	640	1280	768	1024	1024	768
Feed-forward layer dimension	4096	2560	5120	3072	4096	4096	3072
Number of parameters	420M	150M	650M	86M	202M	420M	86M

A.3 Dataset Analysis

Figure 5 is enlarged version of Figure 2(b) and (c).

Sequence Clustering In Section 3.2 and 4.2, we performed clustering of VHH sequences using MMseqs2 v15-6f452 [63][3], released under the GPL-3.0 license. We used the default parameters of easy-cluster, as shown in the following command.

```
mmseqs easy-cluster input.fasta clusterRes tmp --min-seq-id 0.9
```

All input and output files are available at the GitHub repository[4].

Dimensionality Reduction To obtain a 2D representation of the VHH binders as shown in Figure 2(b) and (c), we encoded VHH sequences using Kidera factors [23], which represent the physicochemical properties of amino acids in a 10-dimensional vector, and then converted them into 2D vectors using t-SNE [37]. To run t-SNE, we used sklearn.manifold.TSNE from scikit-learn v1.5.0[5], released under the BSD 3-Clause License, with default parameters, e.g., perplexity=30 and n_iter=1000.

Pairwise Sequence Identities To calculate pairwise sequence identities within AVIDa-SARS-CoV-2 and VHHCorpus-2M, as shown in Figure 3, we used the pairwise2.align.globalxx function of Biopython v1.83 [11][6], released under the BSD 3-Clause License.

A.4 Benchmarks

The code to run the pre-training of VHHBERT and fine-tuning of all baseline models is available at `https://github.com/cognano/AVIDa-SARS-CoV-2`.

A.4.1 Model Implementations

We summarize the model parameters for all baseline models in Table 6.

- **ProtBert** [15] is a BERT-based [13] model pre-trained on 216 million protein sequences in UniRef100 [65]. We used a pre-trained ProtBert[7] released under Academic Free License v3.0 on the Hugging Face Hub.
- **ESM-2** [32] is a protein language model pre-trained on protein sequences in UniRef [65]. During training, sequences are sampled with even weighting across ~43 million UniRef50 training clusters from ~138 million UniRef90 sequences so that over the course of training the model sees ~65 million unique sequences. We used a pre-trained ESM-2 with 150[8] and 650[9] million parameters released under the MIT License on the Hugging Face Hub.
- **AbLang-H** [47] is a RoBERTa-based [33] model pre-trained on 14 million heavy chains of antibodies in the OAS database. We used a pre-trained AbLang-H released in AbLang

[3] MMseqs2: `https://github.com/soedinglab/MMseqs2`

[4] AVIDa-SARS-CoV-2: `https://github.com/cognano/AVIDa-SARS-CoV-2`

[5] scikit-learn: `https://github.com/scikit-learn/scikit-learn`

[6] Biopython: `https://github.com/biopython/biopython`

[7] ProtBert: `https://huggingface.co/Rostlab/prot_bert`

[8] ESM-2 150M: `https://huggingface.co/facebook/esm2_t30_150M_UR50D`

[9] ESM-2 650M: `https://huggingface.co/facebook/esm2_t33_650M_UR50D`

v0.3.1[10], an open source library under the BSD 3-Clause License. Because AbLang-H has a positional embedding layer with a maximum length of 160, we trimmed the first 10 amino acids of the input sequence, which was derived from a phagemid vector rather than a VHH, to accommodate the maximum sequence length of 166 in AVIDa-SARS-CoV-2.

- **AntiBERTa2** [5] is a RoFormer-based [64] model pre-trained using 823.7 million antibody sequences in the OAS and proprietary databases, consisting of 821.2 million unpaired antibody sequences and 2.5 million paired heavy and light chain antibody sequences. We used a pre-trained AntiBERTa2[11] released under a modified version of the Apache 2.0 License[12] on the Hugging Face Hub.
- **AntiBERTa2-CSSP** [5] is a multimodal version of AntiBERTa2. AntiBERTa2-CSSP was trained using 1,554 human antibody structures in SAbDab [14] via contrastive sequence-structure pre-training (CSSP), which amalgamates the representations of antibody sequences and structures in a mutual latent space. We used a pre-trained AntiBERTa2-CSSP[13] released under a modified version of the Apache 2.0 License[14] on the Hugging Face Hub.
- **IgBert** [22] is a model initialized with weights of ProtBert and trained using more than two billion unpaired sequences of light and heavy chains and two million paired sequences in the OAS database. We used a pre-trained IgBert[15] released under the MIT License on the Hugging Face Hub.
- **VHHBERT** is a RoBERTa-based model pre-trained on two million VHH sequences in VHHCorpus-2M. We implemented VHHBERT by using transformers v4.41.1[16] released under the Apache License 2.0. We used the same model parameters as RoBERTa$_{\text{BASE}}$, except that it used positional embeddings with a length of 185 to cover the maximum sequence length of 179 in VHHCorpus-2M. We released the pre-trained VHHBERT[17] under the MIT License on the Hugging Face Hub.

A.4.2 Pre-training Details

As a pre-training corpus for VHHBERT, VHHCorpus-2M was randomly divided into 2,000,000 training sets and 40,988 validation sets. The VHH sequences were tokenized using a vocabulary file[18] consisting of 25 tokens: 20 amino acids and 5 special tokens, resulting in amino acids in the VHH sequence being mapped to different token IDs. Each VHH sequence was padded to the maximum length in each mini-batch during training. We used masked language modeling as the pre-training objective. During pre-training, 15% of the residues from each VHH sequence were randomly selected, and of these, 80% were masked, 10% were randomly changed to another residue, and 10% remained unchanged.

VHHBERT was pre-trained for 312,500 steps, which equates to 20 epochs, with a batch size of 128 on a single machine with one NVIDIA Tesla V100 GPU on the SAKURA cloud[19]. The learning rate was warmed up over the first 5% of the total steps to a peak learning rate of 1e-4 and linearly decayed thereafter. We used the AdamW optimizer [35] with β_1 = 0.90, β_2 = 0.98, ϵ = 1e-6, and a weight decay of 0.01. The training time was approximately three days.

A.4.3 Fine-tuning Details

As a fine-tuning dataset, we used AVIDa-SARS-CoV-2, which was divided by individual, as shown in Table 3. AVIDa-SARS-CoV-2 has the amino acid sequences of VHHs and antigens as input features for binding prediction. To obtain each sequence representation, we used the nine baseline models

[10] AbLang: `https://github.com/oxpig/AbLang`

[11] AntiBERTa2: `https://huggingface.co/alchemab/antiberta2`

[12] AntiBERTa2 License: `https://huggingface.co/alchemab/antiberta2/blob/main/LICENSE.md`

[13] AntiBERTa2-CSSP: `https://huggingface.co/alchemab/antiberta2-cssp`

[14] AntiBERTa2-CSSP License: `https://huggingface.co/alchemab/antiberta2-cssp/blob/main/LICENSE.md`

[15] IgBert: `https://huggingface.co/Exscientia/IgBert`

[16] huggingface/transformers: `https://github.com/huggingface/transformers`

[17] COGNANO/VHHBERT: `https://huggingface.co/COGNANO/VHHBERT`

[18] Vocabulary file: `https://huggingface.co/COGNANO/VHHBERT/blob/main/vocab.txt`

[19] SAKURA cloud: `https://cloud.sakura.ad.jp`

described in Section 5.2 for VHHs and the pre-trained protein language model ESM-2 150M for antigens, respectively. We used the mean of the representations for each amino acid from the last layer in each language model, which is independent of the sequence length, as a sequence representation. Because the maximum length of an antigen sequence is 1328 and ESM-2 150M has a positional embedding layer with a maximum length of 1024, the antigen sequence was divided by 1000 and input into ESM-2 150M separately. The obtained representations were concatenated and averaged to form a sequence representation of the antigen. The sequence representations of the VHHs and antigens obtained from the language models were concatenated and utilized as input to a multi-layer perceptron with one hidden layer of 768 neurons, which was added on top of the two language models as a classification head. Note that we fixed the weights of ESM-2 150M used for antigens and fine-tuned the classification head and the language model used for VHHs to assess the representation capabilities of antibody language models.

We trained the models for 30 epochs with a batch size of 32 on a single machine with one NVIDIA Tesla V100 GPU on the SAKURA cloud. We performed early stopping with a patience of five epochs based on the F1-score on the validation set, which was a randomly sampled 10% from the training set, and selected the model with the best F1-score to evaluate the model performance on the test set. The learning rate was warmed up over the first 5% of the total steps to a peak learning rate of 1e-6 and linearly decayed thereafter. We used the AdamW optimizer with β_1 = 0.90, β_2 = 0.98, ϵ = 1e-6, and a weight decay of 0.01. We conducted five repetitive experiments with different random seeds and report the average results and standard derivation.

https://doi.org/10.52202/079017-3688