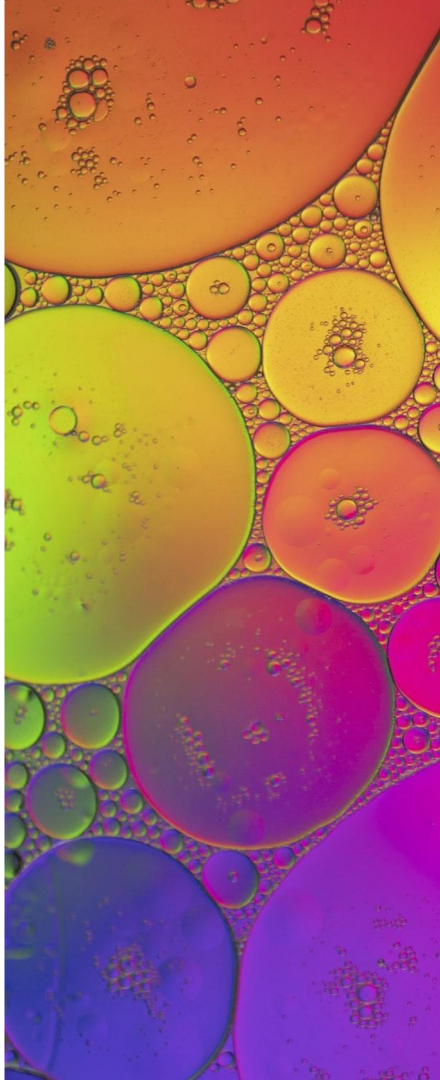
COGNANO SAKURA internet

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

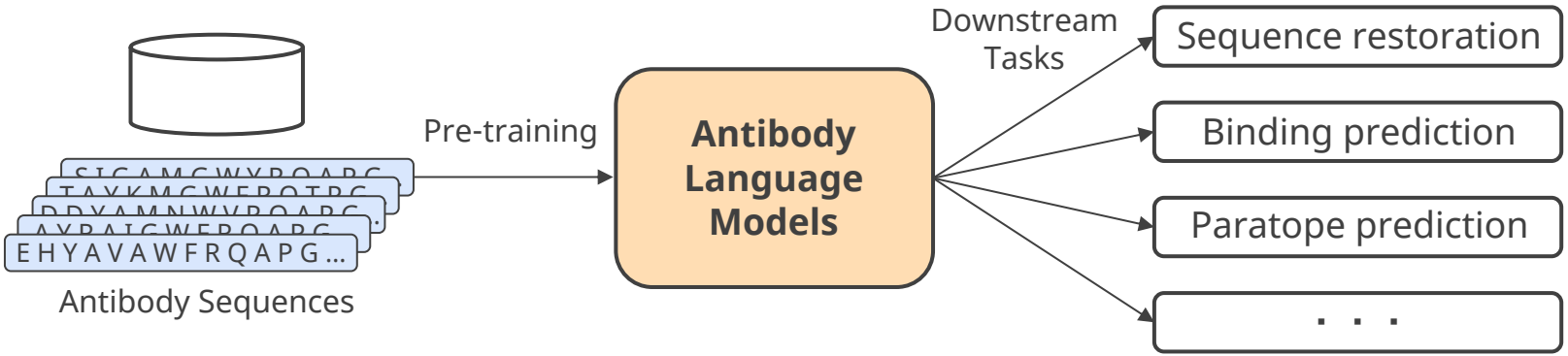
Hirofumi Tsuruta[1,2], Hiroyuki Yamazaki[1,3], Ryota Maeda[1,3], Ryotaro Tamura[1,2], Akihiro Imura[1,3]

[1]COGNANO Inc., [2]SAKURA internet Inc., [3]Biorhodes, Inc.



Background

- **Antibodies** are proteins produced by the immune system to eliminate harmful foreign substances and have become pivotal therapeutic agents for treating human diseases.
- An antibody sequence can be represented as a **string of letters** representing a type of amino acid.
- Recent advances in the **pre-training paradigm** for language models have been increasingly applied to antibody sequences to accelerate antibody discovery.



Pre-trained Antibody Language Models

Characteristics of pre-trained antibody language models.

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

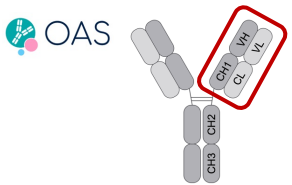
Pre-trained Antibody Language Models

Characteristics of pre-trained antibody language models.

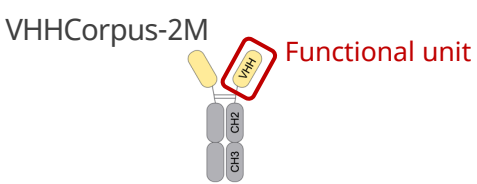
Model	Pre-training			Evaluation		
	Dataset	#Samples	Chain Type	Dataset	#Samples	Task
AntiBERTy	OAS	588M	Heavy, light	HIV-1 donor repertoires	232,593	Evolutionary analysis
AntiBERTa	OAS	72M	Heavy, light	SAbDab	900	Paratope prediction
AbLang-H	OAS	14M	Heavy	OAS	2,000	Sequence restoration
AbLang-L	OAS	0.24M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

Limitations

- Lack of paired heavy and light chain sequences



Conventional antibody



Camelid heavy-chain antibody

Pre-trained Antibody Language Models

Characteristics of pre-trained antibody language models.

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

Limitations

- Small sample size
- Lack of full-length antibody sequences

Overview of the Released Datasets

1. VHHCorpus-2M

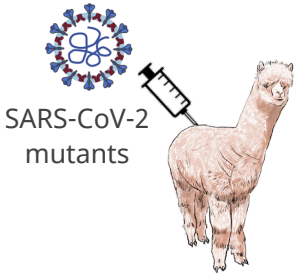
× 5

VHH sequence
K P E D T A V ...
K V D D A A V ...
⋮

Pre-training

2,040,988 samples

2. AVIDa-SARS-CoV-2



× 2

Labeling method※

VHH sequence	Antigen sequence	Label
D R T S W S A ...	M F V F L V L L ...	1
G S R T Y Y A ...	M P M G S L Q ...	0
⋮	⋮	⋮

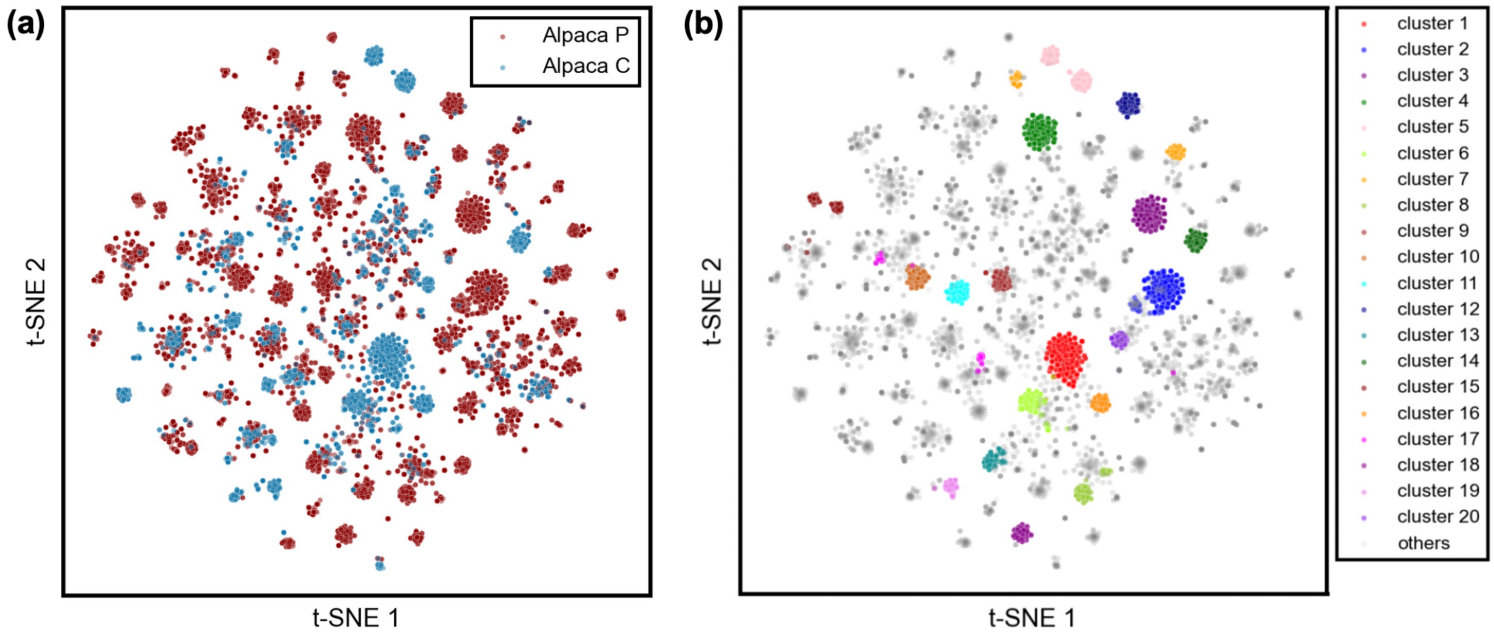
Fine-tuning
Evaluation

77,003 samples

※ Tsuruta, H. et al.: 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)

Individual Differences in SARS-CoV-2-specific VHHs

Two-dimensional representation of binder sequences colored by (a) individuals and (b) clusters with 95% sequence identity. Each plot represents a unique VHH binder.



Using multiple individuals contributes to enhancing the diversity of antigen-specific VHH sequences and provides valuable insights into individual differences in antibody production.

Benchmarks: Task & Models

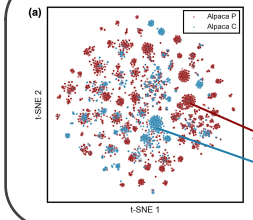
VHHCorpus-2M

Pre-training

Pre-trained Language Models

Fine-tuning using AVIDa-SARS-CoV-2

Binding prediction



Dataset split

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

Baseline Models

	Model	#Parameters	Pre-training Dataset	#Samples
Protein	ProtBert	420M	UniRef	216M
	ESM-2 150M	150M	UniRef	65M
	ESM-2 650M	650M	UniRef	65M
Antibody	AbLang-H	86M	OAS	14M
	AntiBERTa2	202M	OAS, proprietary database	824M
	AntiBERTa2-CSSP	202M	OAS, proprietary database SAbDab	824M 1,554
	IgBert	420M	OAS	2B
	VHHBERT	86M	VHHCorpus-2M	2M
No pre-training	VHHBERT w/o PT	86M	-	-

Benchmarks: Results

Performance comparisons of baseline models for VHH-antigen binding prediction.

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

- Pre-training with antibody sequences, rather than general proteins, contributes to the performance of antibody-specific tasks.
- Additional pre-training of AntiBERTa2-CSSP using human antibody structures contributed to improved performance in predicting VHH-antigen binding.
- AVIDa-SARS-CoV-2 provides valuable benchmarks for evaluating the representation capabilities of antibody language models for binding prediction.

Benchmarks: Results

Performance comparisons of baseline models for VHH-antigen binding prediction.

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

- Pre-training with antibody sequences, rather than general proteins, contributes to the performance of antibody-specific tasks.
- Additional pre-training of AntiBERTa2-CSSP using human antibody structures contributed to improved performance in predicting VHH-antigen binding.
- AVIDa-SARS-CoV-2 provides valuable benchmarks for evaluating the representation capabilities of antibody language models for binding prediction.

Benchmarks: Results

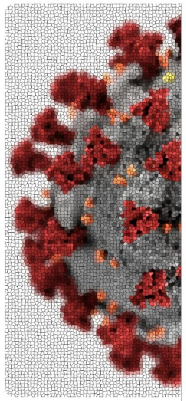
Performance comparisons of baseline models for VHH-antigen binding prediction.

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

- Pre-training with antibody sequences, rather than general proteins, contributes to the performance of antibody-specific tasks.
- Additional pre-training of AntiBERTa2-CSSP using human antibody structures contributed to improved performance in predicting VHH-antigen binding.
- AVIDa-SARS-CoV-2 provides valuable benchmarks for evaluating the representation capabilities of antibody language models for binding prediction.

Thank You!

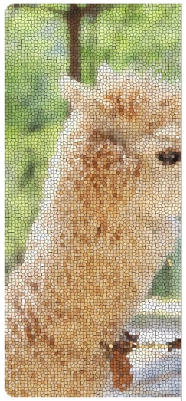
Project Page: https://datasets.cognanous.com



AVIDa-SARS-CoV-2

AVIDa-SARS-CoV-2 is 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 Omicr...

Learn more



VHHCorpus

VHHCorpus is a pre-training corpus with full-length amino acid sequences of variable domain of heavy chain of heavy chain antibody (VHH) collected from alpacas. We currently released VHHCorpus-2M containing over two million unlabeled VHH sequences. VHHCorpus-2M can be used for pre-training of VHH-specific language models.

Learn more

Scan QR Code