Revealing bias in antibody language models through systematic training data processing with OAS-explore

Wiona Sophie Glänzer [1], Sai T. Reddy [1,2], Alexander Yermanos [2,3]

[1] Department of Biosystems Science and Engineering, ETH Zurich, [2] Botnar Institute of Immune Engineering, [3] Center for Translational Immunology, University Medical Center Utrecht

1 Introduction

- Antibody language models aim to unlock insights into immune diversity
- Current training corpora are dominated by a few donors
- OAS-explore is an open-source pipeline to sample more representative training sets through flexible filtering

2 Motivation

- The OAS database with 2.4 billion antibody sequences is commonly used for antibody language model pre-training
- Just 13 individuals from two publications account for 71% of sequences in the database

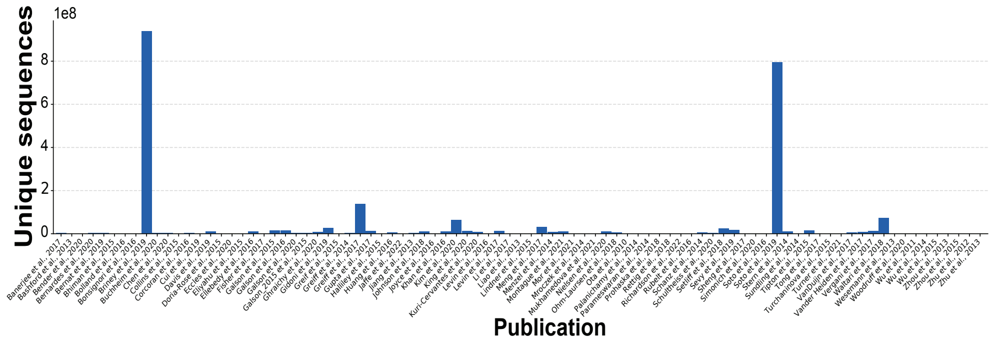
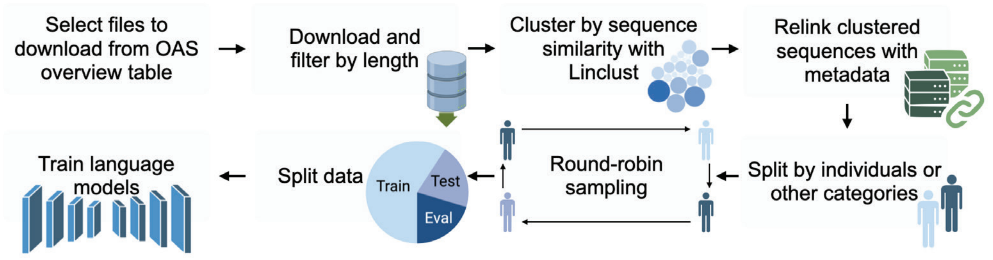


Figure 1: Unique sequences per publication found in the Observed Antibody Space (OAS) database

3 Processing pipeline


Select files to download from OAS overview table
Download and filter by length
Cluster by sequence similarity with Linclust
Relink clustered sequences with metadata
Split by individuals or other categories
Round-robin sampling
Split data
Train
Test
Eval
Train language models

4 Model training

We trained RoBERTa models using a masked language modeling objective on 6 GPUs for 1-6 days per model.

Model name	Number of individuals	Training data size
HIP-1, HIP-2, HIP-3	1 individual per model	3x 30M sequences
Soto-All	3 individuals	90M sequences
OAS-wo-Soto	around 630 individuals	90M sequences

5 Results

We applied similar experimental designs to test generalization across human individuals, heavy and light antibody chains and human versus mouse antibody sequences.

- Models trained only on heavy or only on light chain sequences are not able to fill in the other chain.
- Generalization between human and murine antibody sequences is better, but still very limited.

Generalization to unseen human individuals is poor:

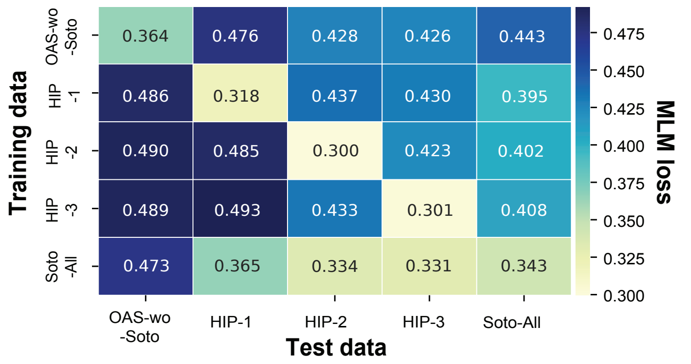


Figure 2: Masked language modeling loss (MLM) of models trained on sequences from 1 (HIP-1, HIP-2, HIP-3), 3 (Soto-All) or 630 (OAS-wo-Soto) individuals on test sets corresponding to every training configuration.

Including more individuals in training data does not improve generalization:

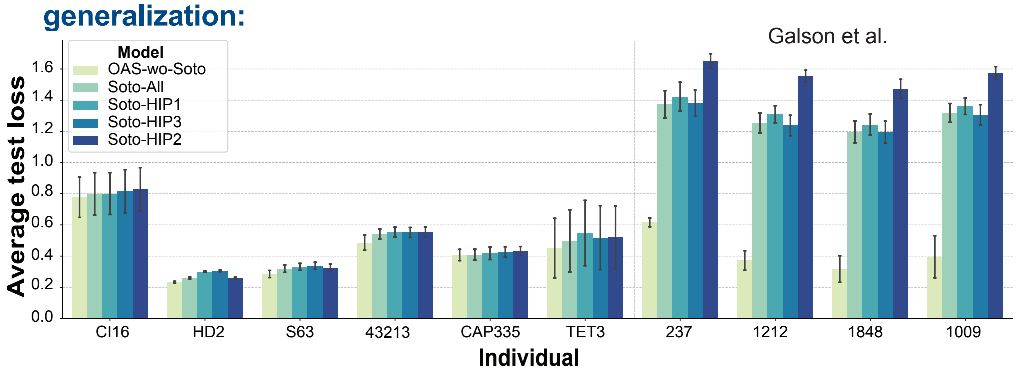


Figure 3: Average MLM loss on sequences from held-out individuals, Subject-237, -1009, -1212, and -1848 are from vaccine studies by the same research group

Batch effects from different experimental protocols strongly impact model performance:

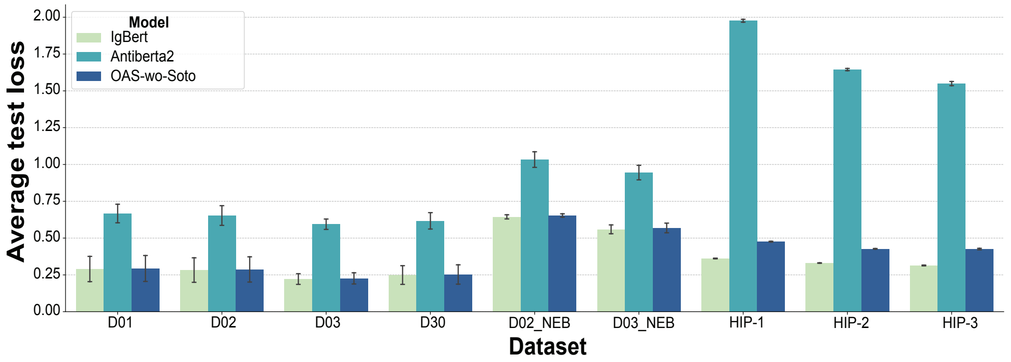


Figure 4: Average MLM loss on antibody sequences from individuals generated with different experimental protocols

5 Conclusion

- Broader donor diversity in training data does not improve generalization of models to unseen human repertoires
- But diverse models are able to compensate publication specific effects
- Even better balancing between individuals, larger training datasets or improved model architectures may be required to achieve robust performance across individuals

Website: Paper: LinkedIn: