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

Wiona Sophie Glänzer
Dep. of Biosystems Science and Engineering
ETH Zürich
4056 Basel, Switzerland
wglaenzer@bsse.ethz.ch

Sai T. Reddy*
Dep. of Biosystems Science and Engineering
ETH Zürich
4056 Basel, Switzerland
sai.reddy@bsse.ethz.ch

Alexander Yermanos*
Center for Translational Immunology
University Medical Center Utrecht
3584 CX Utrecht, The Netherlands
alex.yermanos@immune.engineering

Abstract

Antibody language models (LMs) trained on immune receptor sequences have been applied to diverse immunological tasks such as humanization and prediction of antigen specificity. While promising, these models are often trained on datasets with limited donor diversity, raising concerns that biases in the training data may hinder their generalizability. To quantify the impact of biased training data, we introduce an open-source processing pipeline for the 2.4 billion unpaired antibody sequences in the Observed Antibody Space (OAS) database, enabling customizable filtering and balanced sampling by donor, species, chain type and other metadata. Analysis of OAS revealed that 13 individuals contribute over 70% of human antibody sequences. Using our pipeline, we trained 17 RoBERTa antibody LMs on datasets of different compositions. Models failed to generalize across chain types and showed limited transfer between human and mouse repertoires. Both individual- and batch-specific effects influenced model performance, and expanding donor diversity did not improve generalization to unseen individuals from unseen publications.

1 Introduction

Protein language models (LMs), which treat amino acid sequences as a form of biological language, can learn rich, contextual representations of protein structure and function from unlabeled sequence data [1]. Among proteins, antibodies are particularly compelling targets for such modeling due to their important role in the immune system and growing success as therapeutics [2], but their vast diversity in sequences and structures also presents a unique challenge [3]. Specialized antibody LMs have emerged to support tasks such as paratope prediction, structure inference, and affinity optimization, and these models rely heavily on large-scale datasets for pretraining [4].
The Observed Antibody Space (OAS) database is the largest public collection of antibody sequences and the main source of training data for all public antibody LMs [5]. Despite its central role, there has been little scrutiny of its biases, no reproducible pipeline for preparing training data from the database exists, and the training sets of existing models have not been published. This lack of transparency

*Botnar Institute of Immune Engineering, 4056 Basel, Switzerland

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences .

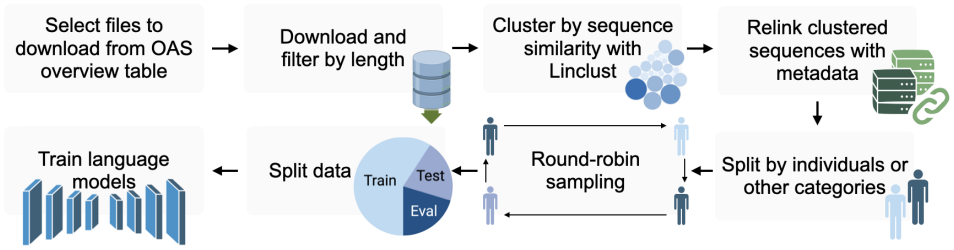

Figure 1: Processing pipeline for antibody sequence data from OAS

hampers our ability to assess the true capabilities and limitations of antibody LMs and to design better training strategies. In this work, we introduce an open pipeline for processing antibody sequence data from OAS and investigate how dataset composition and biases influence model performance. All code, models and datasets are available on GitHub and Hugging Face.

2 A processing pipeline for antibody language model training data from OAS

2.1 Steps of the processing pipeline

To explore how data composition impacts the performance of antibody-specific LMs, we created a Snakemake pipeline, OAS-explore, which unifies common steps for preparing OAS data for antibody LM pretraining (Figure 1). First, V(D)J sequences are downloaded and a length filtering (as in AntiBERTa [6]) is applied. Next, sequences are clustered by similarity with Linclust [7] and afterwards mapped back to their corresponding metadata to support additional filtering and analysis of performance on metadata-defined subsets of test data. We add a new processing step that partitions data by individual donors and allows us to balance the composition of the training data by round-robin sampling. Finally, data is split into training, test and evaluation sets, sequences are tokenized and LMs trained.

2.2 Analysis of the composition of OAS

The OAS database contains approximately 2.4 billion predominantly human unpaired antibody sequences, of which 71% come from just two studies: Briney et al. 2019 [8] (10 donors) and Soto et al. 2019 [9] (3 donors); see Supplementary Figure S1. Many sequences from Briney et al. fail framework region 1 length filtering, so we primarily use data from Soto et al. for experiments. Although more than 630 individuals appear in OAS, most are represented by only a few sequences. This biased donor distribution is also reflected in the training datasets of most antibody LMs ([6], [10], [11], [12], [13]). We used our pipeline to analyze the effect of this bias on model performance and generalization.

3 Relevance of chain and species in the training data

Previously, antibody LMs have been trained either only on human data (e.g., Sapiens [13]) or mixed data from all species included in OAS (e.g., AntiBERTy [11], AbLang [10]). Antibodies consist of two types of amino-acid chains. Some models are trained on datasets that pool both heavy- (IGH) and light-chain (IGK or IGL) sequences (e.g., AntiBERTa [6]), whereas others train separate models for each chain type (e.g., Sapiens [13]). Fine-tuning of models pretrained on general protein datasets on antibody data has also been used (e.g., IgBert [12]). The effects of these design choices are not well understood.

3.1 Model performance for mixed and separate training

We trained nine RoBERTa models, each on 1M sequences for 10 epochs. For both human and mouse data, we fit light-only, heavy-only, and 50:50 light-heavy models, and we also trained three joint human–mouse counterparts. Due to limited availability of mouse light chain sequences, 250K unique

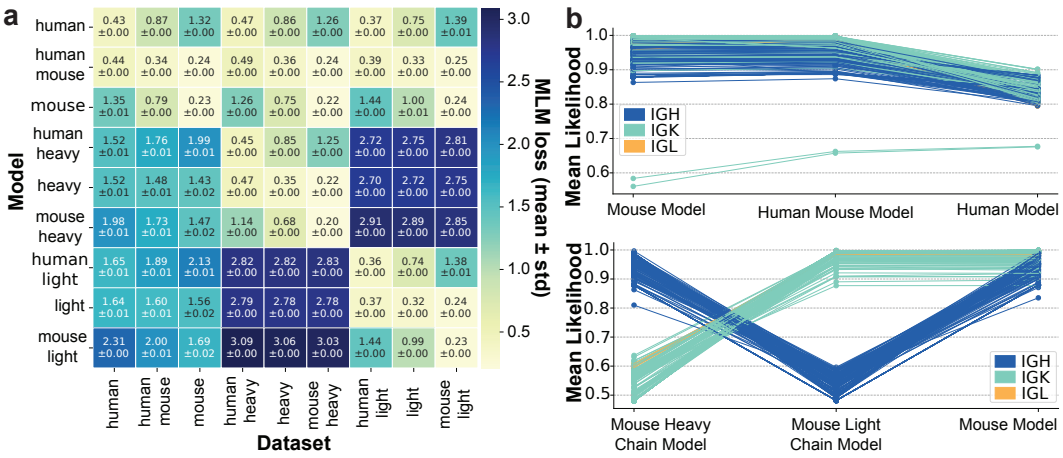

Figure 2: Comparison of models trained on mouse, human and mixed datasets **a.** MLM loss on test sets of 100k sequences for models trained with varying species and chain composition. Y-axis: training data origin; x-axis: test data origin. **b.** Average AA-likelihoods for sequences representing 5% of a mouse bone marrow repertoire from [16]. Sequences are colored by chain type; x-axis specifies model training data composition.

sequences were repeated four times (Supplementary Table T2). Each model was evaluated on all test sets to assess generalization across antibody chains and species (Figure 2a).
Models achieved the lowest MLM loss when evaluated on test sets that matched their training data composition. Cross-chain generalization was negligible: the loss of models trained on heavy chain sequences, when tested on light chain sequences, was almost as high as the initial loss of randomly initialized models during training (Figure 2a). Cross-species generalization was better but still limited. Interestingly, models trained on mixed datasets performed nearly as well on individual test sets (e.g., light or heavy chains) as models trained exclusively on the corresponding data type (Figure 2a). When adding just 1% mouse sequences to otherwise human training data, performance on the mouse test set improved substantially (MLM loss of 0.41 vs. 1.3 for 0% mouse). This suggests that the model can transfer some knowledge about antibody sequences from human to mouse sequences, even with minimal exposure during training (S2).

3.2 Species identity drives sequence likelihoods

Protein language model-derived likelihoods have been used in antibody engineering to suggest mutations that increase binding affinity [14] and for antibody humanization [15]. We looked at how the average likelihood

$$\bar{L}(s) = \frac{1}{|s|} \sum_{i=1}^{|s|} p_\theta(s_i \mid s_{\setminus i})$$

of a sequence s is influenced by training data composition. Average likelihoods of mouse antibody sequences are significantly higher when computed using a model trained on mouse or mixed-species data compared to a model trained solely on human sequences. A similar pattern occurs when comparing likelihoods of heavy and light chain models (Figure 3a).

4 Model performance depends on individual and batch of origin

Given the over-representation of a small number of individuals in the OAS database, we investigated whether antibody LMs learn a universal antibody "language" or mainly pick up individual-specific features by comparing models trained on one versus many individuals.

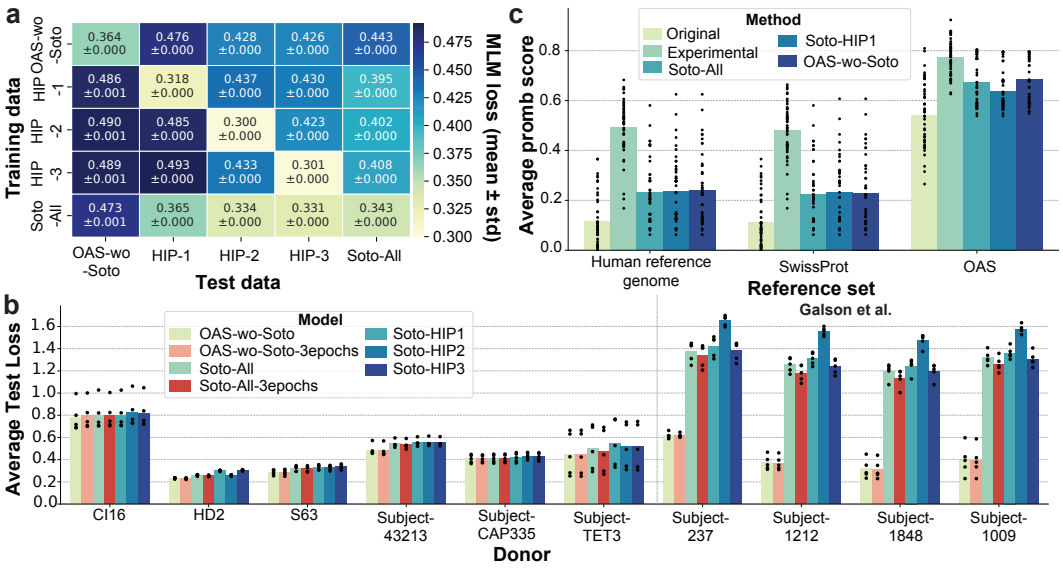

Figure 3: Impact of training data diversity on model performance **a.** MLM loss of models trained on sequences from 1 individual (HIP-1, HIP-2, HIP-3), 3 individuals (Soto-All) or 630 individuals (OAS-wo-Soto), evaluated on test sets corresponding to each training configuration. **b.** Average MLM loss on sequences from held-out individuals. Subject-237, -1009, -1212, and -1848 are from vaccine studies by the same research group. **c.** Average humanization scores across 25 antibodies

4.1 Generalization across individuals

We trained three single-donor models (HIP-1/2/3) using data from Soto et al. [9], a combined Soto-All model, and a more balanced OAS-wo-Soto model on a dataset created via round-robin sampling excluding individuals HIP-1/2/3 (A.2). Although some imbalance remains, the resulting 90 million sequence dataset is markedly more balanced than our comparison sets (S3) and the training data in previous antibody LM studies.

Each model performs best on test data matching its training data composition and generalization to unseen individuals is limited based on MLM loss (Figure 3a). The balanced OAS-wo-Soto model performs similarly to single-donor models (HIP-1/2/3) on unseen individuals, indicating that increasing the number of individuals in the training data alone does not improve generalization. Loss differences between the HIP-1 and HIP-2/3 test sets are due to a higher proportion of heavy chains in the HIP-1 set.

On ten additional held-out individuals, performances vary considerably. For individuals HD2 and S63, all models, including single-donor ones, achieve losses close to those on matched training-test splits (Figure 3b). Conversely, for four individuals from vaccination studies by the same research group, the models trained on few individuals (e.g., HIP-1, Soto-All) perform significantly worse than the balanced OAS-wo-Soto model. This model may be able to compensate for batch-specific effects because it was trained on other individuals from those studies. Continuing training for two additional epochs did not improve performance on unseen individuals (Figure 3b), and our models show similar or better performance than previously published models IgBert and AntiBERTa-2 (S4).

4.2 Humanization of antibody sequences

When LMs are used for humanization, we assume that they have implicitly learned what makes an antibody human, but training data bias might lead to models learning a skewed representation of "humanness". We tested the humanization capabilities of our models on 25 antibody sequences with known experimental humanizations [15]. We used an iterative mutation procedure (A.5) and the promb scoring system [17], which measures the proportion of 9-mer peptides in a sequence that appear in a reference database. With OAS as the reference, model-based humanizations scored nearly as well as experimental ones; with other references, they scored significantly worse (Figure 3c). We

observed no differences in scores between our models for any of the references, but more detailed evaluation tools might be needed to uncover effects of training data diversity on humanization.

5 Discussion

Using OAS-explore, we curated OAS subsets to train 17 RoBERTa antibody LMs and found that models struggle to generalize to new individuals and unseen batches. Future work should aim to disentangle these two factors and develop preprocessing strategies to mitigate biases. Limitations of our work include small training sets for species comparisons (Section 3) due to scarcity of mouse sequences. Larger training sets might improve generalization. For humans (Section 4), round-robin sampling balanced individuals better than random selection, yet some donors remained over-represented. Data from more individuals may be required to achieve generalizability of models. We did not retrain models, and standard deviations reflect five test-set splits. To facilitate follow-up work, we release all datasets together with an easy-to-use pipeline, lowering barriers to systematic tests of training-data composition and fostering open, reproducible development of antibody LMs.

Acknowledgments and disclosures

We would like to thank the Euler cluster at ETH Zurich for providing computational resources for model training and testing.

Funding: This research received no external funding.

Competing interests: S.T.R. is a co-founder and holds shares of Engimmune Therapeutics AG and Encelta and Fy Cappa Biologics. S.T.R. may hold shares of Alloy Therapeutics. S.T.R. is on the scientific advisory board of Engimmune Therapeutics, Alloy Therapeutics, Encelta and Fy Cappa Biologics. S.T.R. is a member of the board of directors for Engimmune Therapeutics and GlycoEra.

Data and materials availability: All code, models and datasets are available at github.com/WionaGlaenzer/OAS-explore and huggingface.co/collections/WionaGlaenzer/oas-explore-68da7c225ba55eb5c895b930.

References

[1] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, L. C. Zitnick, J. Ma, and R. Fergus. 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. URL https://doi.org/10.1073/pnas.2016239118.

[2] A. C. Chan, G. D. Martyn, and P. J Carter. Fifty years of monoclonals: The past, present and future of antibody therapeutics. *Nat Rev Immunol*, 25(8):497–510, 2025. URL https://doi.org/10.1038/s41577-025-01207-9.

[3] V. Greiff, E. Miho, U. Menzel, and S. T. Reddy. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.*, 36:738–749, 2015. URL https://doi.org/10.1016/j.it.2015.09.006.

[4] D. Wang, F. Ye, and H. Zhou. On pre-trained language models for antibody. *arXiv*, 2023. URL https://doi.org/10.48550/arXiv.2301.12112.

[5] T. H. Olsen, F. Boyles, and Deane C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 2021. URL https://doi.org/10.1002/pro.4205.

[6] J. Leem, L. S. Mitchell, J. H. R. Farmery, J. Barton, and J. D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022. URL https://doi.org/10.1016/j.patter.2022.100513.

[7] M. Steinegger and J. Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(2542), 2018. URL https://doi.org/10.1038/s41467-018-04964-5.

[8] B. Briney, A. Inderbitzin, C. Joyce, and D. Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566:393–397, 2019. URL https://doi.org/10.1038/s41586-019-0879-y.

[9] C. Soto, R. G. Bombardi, A. Branchizio, N. Kose, M. Pranathi, A. M. Sevy, R. S. Sinkovits, P. Gilchuk, J. A. Finn, and J. E. Crowe. High frequency of shared clonotypes in human b cell receptor repertoires. *Nature*, 566:398–402, 2019. URL https://doi.org/10.1038/s41586-019-0934-8.

[10] T. H. Olsen, I. H. Moal, and C. M. Deane. Ablang: An antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022. URL https://doi.org/10.1093/bioadv/vbac046.

[11] J. A. Ruffolo, J. J. Gray, and J. Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv*, 2021. URL https://doi.org/10.48550/arXiv.2112.07782.

[12] H. Kenlay, F. A. Dreyer, A. Kovaltsuk, D. Miketa, D. Pires, and C. M. Deane. Large scale paired antibody language models. *PLOS Computational Biology*, 20(12):e1012646, 2024. URL https://doi.org/10.1371/journal.pcbi.1012646.

[13] D. Prihoda, J. Maamary, A. Waight, V. Juan, L. Fayadat-Dilman, D. Svozil, and D. A. Bitton. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. URL https://doi.org/10.1080/19420862.2021.2020203.

[14] B. L. Hie, V.R. Shanker, D. Xu, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, J. E. Wu, W.and Pak, and P. S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, 2024. URL https://doi.org/10.1038/s41587-023-01763-2.

[15] N. L. Hepler, A. J. Hill, D .B. Jaffe, M. C. Gibbons, K. A. Pfeiffer, D. M. Hilton, M. Freeman, and W. J. McDonnell. Better antibodies engineered with a glimpse of human data. *bioarXiv*, 2025. URL https://doi.org/10.1101/2025.06.08.658113.

[16] A. Agrafiotis, D. Neumeier, K. Hong, T. Chowdhury, R. Ehling, R. Kuhn, I. Sandu, V. Kreiner, T. S. Cotet, D. Shlesinger, D. Laslo, S. Anzböck, D. Starkie, D. J. Lightwood, A. Oxenius, S. T. Reddy, and A. Yermanos. Generation of a single-cell b cell atlas of antibody repertoires and transcriptomes to identify signatures associated with antigen specificity. *iScience*, 26(3):106055, 2023. URL https://doi.org/10.1016/j.isci.2023.106055.

[17] D. Prihoda. promb: Protein humanness evaluation toolkit. *GitHub*. URL https://github.com/MSDLLCpapers/promb.

A Technical Appendices and Supplementary Material

A.1 Data processing

Sequences were downloaded from OAS on 29.04.2025. For download and data processing we used our custom pipeline OAS-explore. We activated length filtering as in Leem et al. [6], which uses the following criteria: at least 20 residues in the FR1, at least 10 residues in the FR4, 5-12 residues in the CDR1, 1-10 in the CDR2, and 5-38 residues in the CDR3. The criterion for FR1 is the strongest filter, excluding the highest number of sequences. We also specified filters for species, publication and chain in OAS-explore depending on the desired composition of each dataset.
For similarity clustering with Linclust we used a coverage and a similarity of 0.9. As the sampling scheme we selected "random" for most datasets and "round-robin" for the OAS-wo-Soto dataset. We used "numbers" as the "split_mode" and specified the exact number of sequences in train/test/eval sets in the configuration file. An overview of the datasets used for our experiments can be found in Supplementary Table T2.

A.2 Model training

RoBERTa models were trained using a masked language modeling objective with the HuggingFace Transformers library and training code modified from Leem et al. [6]. All models had 12 hidden layers and 12 attention heads with a hidden size of 768. The vocabulary size is 25 with single-amino-acid tokenization and the maximum sequence length is 150 tokens. 15% of tokens were masked. The per-device batch size was 96. We used a linear learning rate schedule with a warm-up period and a peak learning rate of 0.0001. We used the Adam optimizer and a weight decay of 0.01. Hyperparameters were chosen in accordance with Leem et al. [6].
Training was conducted on 6 NVIDIA GeForce RTX 2080 Ti GPUs on an internal cluster. For models used in Section 3, training took up to 2 hours per model. For models used in Section 4, training took between 2 and 12 days. Details can be found in Supplementary Table T1.

A.3 Evaluation with MLM loss

Masked language modeling loss was extracted from transformers.Trainer() for both our and public models. For Figure 2a and Figure S2, test sets contained 100,000 sequences each. For Figure 3a, the test set size was 1.6 million sequences for single-individual models (5%) and 2.5 million sequences for OAS-wo-Soto (3%). These sizes were chosen based on data availability, while keeping training datasets the same size. The Soto-All test set combines all single-individual test sets. All standard deviations were calculated on 5 test set splits using pandas std with $n - 1$. In Figure 3b, dots represent MLM loss on single splits. The ten individuals excluded from all training datasets were selected among individuals with 10,000 to 100,000 sequences present in OAS such that several publications were represented.

A.4 Likelihood calculations

To show the effects of training data on amino acid likelihoods, we randomly sampled 5% from a 7,000-sequences mouse repertoire from [16]. We chose this sample to use sequences not included in the OAS and thus excluded from training. We sampled sequences from a single mouse repertoire to remove subject of origin as a confounding factor. We calculated likelihoods for each sequence position separately and then averaged across the sequence to get a measure that is independent of sequence length.

A.5 Antibody humanization

Humanization is a process by which an antibody drug candidate's sequence, which might have been developed using animal experiments, is modified to adapt it for use in humans. The goal is to keep the functionality of the antibody but reduce the chance that it is recognized as a foreign object by the human immune system by using an amino acid composition similar to that found in natural human antibodies [13].

We employed an iterative humanization procedure that has already been used in previous publications for LM-based humanization of antibody sequences: In each iteration, the model identifies positions where alternative amino acids have higher log-likelihoods than the original ones. The alternative amino acid with the highest log-likelihood is mutated, and the updated sequence is passed to the model again. This process is repeated until no further substitutions are suggested [15].

We calculated the likelihoods at each position by passing the complete sequence to the RobertaForMaskedLM model only masking one position at a time. promb scores were calculated using 9-mers for all references. In Figure 3c, dots represent promb scores of single antibody sequences.

A.6 Supplementary figures

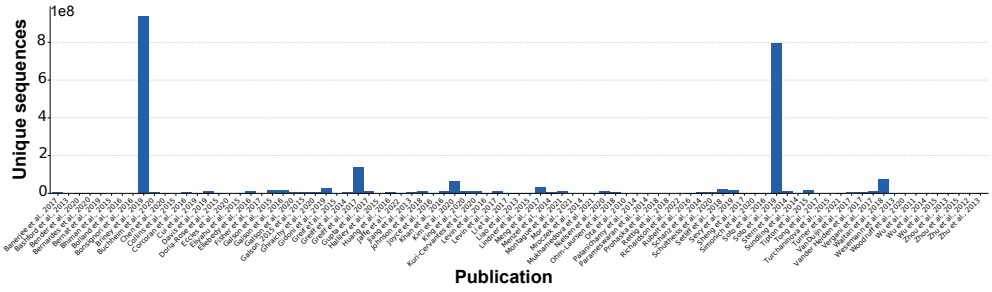

Figure S1: Unique sequences per publication contained in OAS

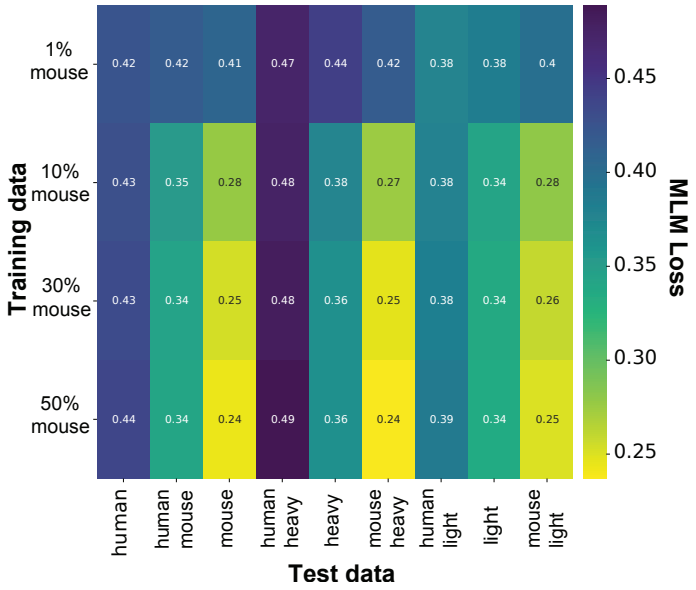

Figure S2: MLM loss with varying percentages of mouse data used in training

Table T1: Models

Model name	Epochs	Training duration
HIP-1	1	20.6 hours
HIP-2	1	26.0 hours
HIP-3	1	27.3 hours
Soto-All	1	3.1 days
OAS-wo-Soto	1	2.8 days
Soto-All-3epochs	3	9.5 days
OAS-wo-Soto-3epochs	3	12 days
human/mouse-models	10	1.7 - 2 hours

Table T2: Datasets

Name	Description	Size
HIP-1/2/3	1 individual from Soto et al. per dataset	3 x 30M
Soto-All	Combination of HIP-1, -2 and -3 datasets	90M
OAS-wo-Soto	Sampled from all individuals in OAS without Soto et al.	90M
human-light	Randomly sampled human light chain sequences	1M
human-heavy	Randomly sampled human heavy chain sequences	1M
human	1/2 of the human-light and 1/2 of the human-heavy dataset	1M
mouse-heavy	Randomly sampled mouse heavy chain sequences	1M
mouse-light	250k mouse light chain sequences	4 x 250k
mouse	1/2 of the mouse-heavy and 2 x 1/4 of the mouse-light dataset	1M
light	1/2 of the human-light and 1/2 of the mouse-light dataset	1M
heavy	1/2 of the human-heavy and 1/2 of the mouse-heavy dataset	1M
human-mouse	1/2 of the human and 1/2 of the mouse dataset, heavy:light 50:50	1M
1% mouse	99% from human and 1% from mouse dataset, heavy:light 50:50	1M
10% mouse	90% from human and 10% from mouse dataset, heavy:light 50:50	1M
30% mouse	70% from human and 30% from mouse dataset, heavy:light 50:50	1M

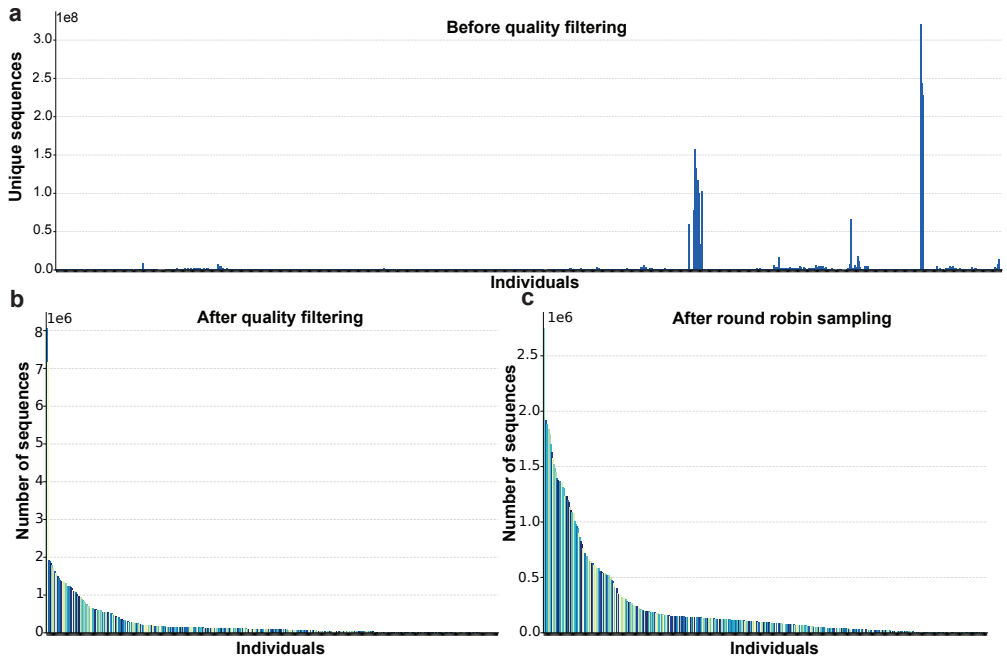

Figure S3: Distributions of unique sequences per individual. **a.** Before quality filtering. **b.** After quality filtering and excluding Soto et al. **c.** After quality filtering, excluding Soto et al. and round-robin sampling; training data for OAS-wo-Soto.

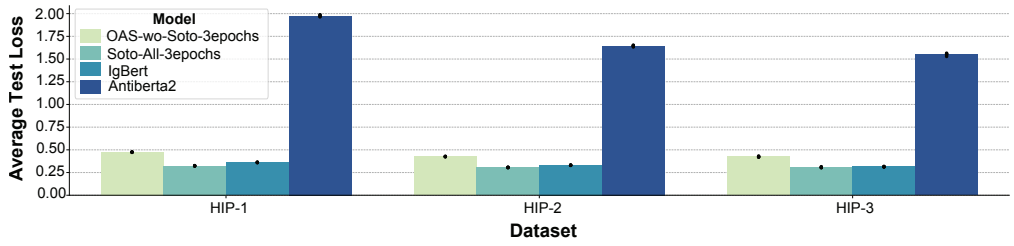

Figure S4: Comparison of MLM loss with public models on test sets of 100,000 sequences from individuals from Soto et al.