

---

# Deciphering antibody affinity maturation with language models and weakly supervised learning

---

**Jeffrey A. Ruffolo**

Johns Hopkins University  
Baltimore, MD 21218  
jruffolo@jhu.edu

**Jeffrey J. Gray**

Johns Hopkins University  
Baltimore, MD 21218  
jgray@jhu.edu

**Jeremias Sulam**

Johns Hopkins University  
Baltimore, MD  
jsulam@jhu.edu

## Abstract

In response to pathogens, the adaptive immune system generates specific antibodies that bind and neutralize foreign antigens. Understanding the composition of an individual’s immune repertoire can provide insights into this process and reveal potential therapeutic antibodies. In this work, we explore the application of antibody-specific language models to aid understanding of immune repertoires. We introduce AntiBERTy, a language model trained on 558M natural antibody sequences. We find that within repertoires, our model clusters antibodies into trajectories resembling affinity maturation. Importantly, we show that models trained to predict highly redundant sequences under a multiple instance learning framework identify key binding residues in the process. With further development, the methods presented here will provide new insights into antigen binding from repertoire sequences alone.

## 1 Introduction

The adaptive immune system is capable of generating robust responses to foreign pathogens. This robustness is provided in part by the immense diversity of antibodies that can be generated. Diversity is initially introduced to antibody sequences through V(D)J gene recombination. As the immune response progresses, antibodies capable of effective neutralization are developed through an antigen-driven process called affinity maturation. In this process, B-cells producing antibodies with high antigen affinity are selectively expanded then mutated to give rise to successive generations of antibodies.

Immune repertoire samples provide a snapshot of an individual’s antibody sequence population. Typically, donors provide samples of B-cells from blood or lymph, and the antibodies produced by these B-cells are identified via next-generation sequencing [5]. During an immune response, as many as half of the antibodies within the repertoire may exhibit antigen affinity [14]. Among these binding antibodies, the most frequently occurring sequences tend to be effective binders [15]. However, beyond the small subset of highly expanded sequences, redundancy is a poor indicator of binding capability [14].

Models adopted from natural language processing and trained on massive sets of protein sequences have been shown repeatedly to learn rich representations of protein sequences [4, 16]. Such models have been used for mutational variant prediction [13], to generate embeddings for structure prediction [1], and even to study evolution within protein families [6]. However, natural proteins evolve under many selective pressures, while antibodies are selected for binding to a particular antigen. As such, models trained on all proteins may be poorly suited for capturing specific features of antibody sequence evolution. In this work, we explore the use of an antibody-specific language model for understanding affinity maturation within immune repertoires.

## 2 Methods

### 2.1 Antibody encoder model

To learn antibody-specific representations, we trained a transformer encoder model [2] on 558M non-redundant sequences from the Observed Antibody Space [11] (Appendix A.1). The model, which we call AntiBERTy, is based on the BERT architecture [2] implementation from Huggingface [19]. We trained using the masked language modeling (MLM) objective, as originally proposed by Devlin et al. [2]. Specific model parameters and training loss plots are provided in Appendix A.2

### 2.2 Evolutionary analysis of immune repertoires

Inspired by recent work in modeling protein evolution with language models [6], we began by looking for global trends within individual repertoires. We considered immune repertoire samples from four donors who developed neutralizing antibodies against HIV-1 [22, 21] (Appendix A.3). Specifically, these donors developed antibodies belonging to the VRC01 class, which bind to the HIV-1 spike glycoprotein120 (gp120) through a  $V_H$ -gene mediated paratope [20, 21]. For each repertoire, we created a  $k$ -nearest-neighbor graph using embeddings from AntiBERTy as described by Hie et al [6]. We additionally plotted evo-velocity scores (Appendix A.4).

### 2.3 Multiple instance learning on sequence embeddings

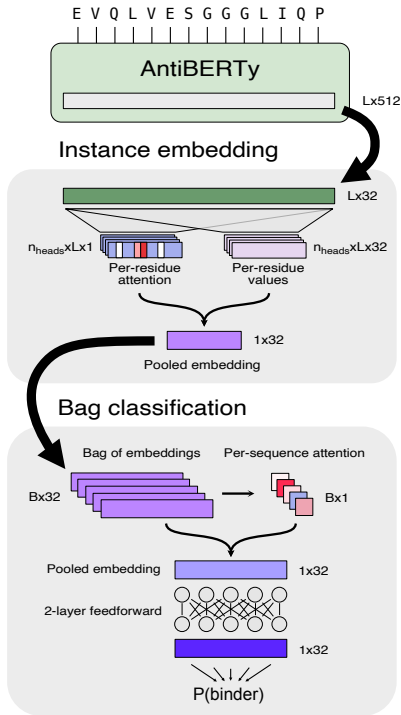


Figure 1: Diagram of MIL model for predicting whether a bag of sequences contains a highly redundant instance.

extract the final hidden representation from AntiBERTy and reduce the dimensionality using a linear transformation, resulting in a variable-length embedding  $H = (\mathbf{h}_1, \dots, \mathbf{h}_L)$  with  $\mathbf{h}_i \in \mathbb{R}^{d_{\text{emb}}}$ . Then, we use a multi-head attention [18] pooling layer to reduce the variable-length embedding to a fixed size vector, described below for one attention head:

$$a_i^{\text{emb}} = \frac{\exp(\mathbf{w}_{\text{emb}}^T (\mathbf{Q}\mathbf{h}_i^T \odot \mathbf{K}\mathbf{h}_i^T))}{\sum_j \exp(\mathbf{w}_{\text{emb}}^T (\mathbf{Q}\mathbf{h}_j^T \odot \mathbf{K}\mathbf{h}_j^T))} \quad (1)$$

in order to obtain the final fixed-size instance embedding  $\mathbf{x} = \sum_i^L a_i^{\text{emb}} \mathbf{V} \mathbf{h}_i^T$ . In the above,  $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{d_{\text{emb}} \times d_{\text{head}}}$  and  $\mathbf{w}_{\text{emb}} \in \mathbb{R}^{d_{\text{head}} \times 1}$  are learnable parameters, and  $\mathbf{a}^{\text{emb}} \in \mathbb{R}^L$  is a per-residue attention score. In practice, we use four attention heads and concatenate the fixed-size embeddings. Each sequence in a bag is passed individually through the instance embedding module and the embeddings are collected to form the set  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_B\}$ . We adopt the gated attention mechanism from Ilse et al [7] to learn a MIL pooling operation:

$$a_i^{\text{bag}} = \frac{\exp(\mathbf{w}_{\text{bag}}^T (\text{sigm}(\mathbf{V} \mathbf{h}_i^T) \odot \text{sigm}(\mathbf{U} \mathbf{h}_i^T)))}{\sum_j^B \exp(\mathbf{w}_{\text{bag}}^T (\text{sigm}(\mathbf{V} \mathbf{h}_j^T) \odot \text{sigm}(\mathbf{U} \mathbf{h}_j^T)))}, \quad (2)$$

finally obtaining

$$\mathbf{z} = \sum_i^B a_i^{\text{bag}} \mathbf{h}_i, \quad (3)$$

where  $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{d_{\text{attn}}}$  and  $\mathbf{w}_{\text{bag}} \in \mathbb{R}^{d_{\text{attn}} \times 1}$  are learnable parameters,  $\mathbf{a}^{\text{bag}} \in \mathbb{R}^B$  is a per-instance attention score, and  $\mathbf{z} \in \mathbb{R}^{d_{\text{emb}}}$  is a fixed-size bag embedding. Finally, the bag classification is predicted with a two-layer feed-forward network followed by the logistic function.

Specific model parameters are provided in Appendix A.5. We use cross-entropy loss to train the model for 20 epochs, where an epoch is defined as the total repertoire size divided by the bag size. During training, we sample positive and negative bags with equal frequency.

### 3 Results

#### 3.1 Language model reveals trajectories within repertoire

We applied the evolutionary analysis to repertoire samples from four donors who produced VRC01 class antibodies. First, we visualized the KNN graph in two-dimensional UMAP [12] embedding and annotated each sequence with the distance from germline (Fig. 2A). For each donor, we observe continuous trajectories between germline sequences and highly mutated derivatives, consistent with the process of affinity maturation. Next, we combined the repertoires from all donors into a single set and repeated the analysis (Fig. 2B). The combined sequence graph displays significant overlap between the antibody sequences from donors RU3, IAVI57, and IAVI74, consistent with previous findings that VRC01 antibodies share ontogenies [22]. In contrast, the repertoire from donor

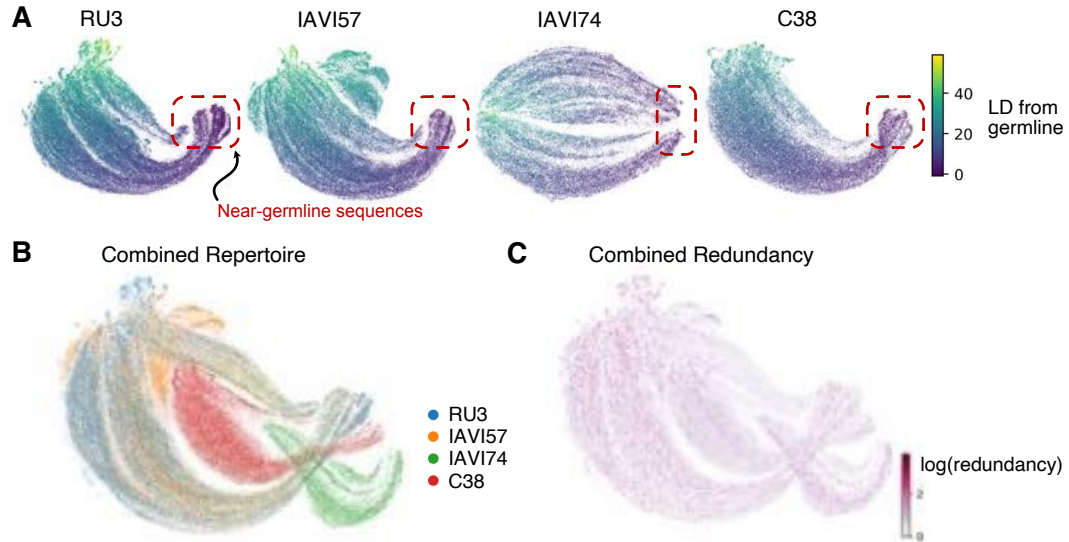


Figure 2: Evolutionary analysis of immune repertoires. (A) UMAP embedding of repertoire sequences annotated with Levenshtein distance (LD) from germline for four donors. (B) UMAP embedding of four-donor combined repertoire. (C) Combined repertoire annotated with sequence redundancy.

C38 overlaps little with other donors, likely due to maturation from an alternative set of germline sequences. Finally, we show the redundancy of each sequence within the combined repertoire (Fig. 2C). We observe a relatively uniform distribution of redundancy throughout the embedded space. This uniformity is likely a reflection of iterative nature of affinity maturation, by which sequences are developed through many rounds of clonal expansion and diversification.

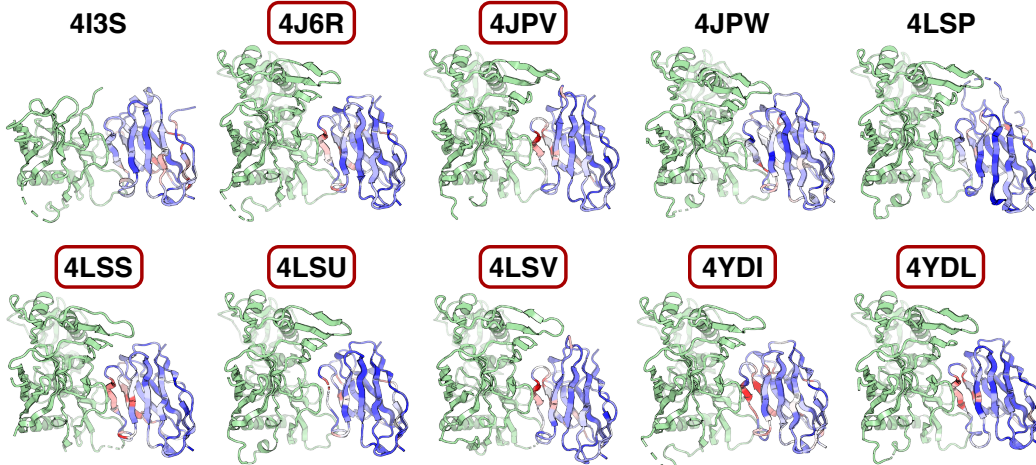


Figure 3: MIL model attention reveals paratope. Ten VRC01 antibodies annotated with residue-level attention from MIL model in complex with gp120 antigen (green). Attention values increase from blue to red. For seven antibodies (indicated with red boxes), attention localizes to paratope residues.

### 3.2 MIL model identifies VRC01 paratope

MIL models were trained for each individual repertoire, as well as the combined repertoire, as described above. For each dataset, the model learned to effectively identify bags containing high-redundancy sequences (Appendix A.6). To verify that the model had learned properties of VRC01 antibodies, we collected a set of ten VRC01 antibodies crystallized in complex with the gp120 antigen [8, 10, 21, 22]. We created single-instance bags for each sequence and confirmed that the MIL model successfully produced positive predictions. Next, we investigated whether the residues contributing to predictions were consistent with the binding mode of VRC01 antibodies. For each complex, we annotated the antibody structure with the attention  $a^{\text{emb}}$  from each of four heads of the instance embedding module (Fig. 3). For seven of the ten sequences, the second attention head showed remarkable localization to the H2 loop and the following beta strand, consistent with the VRC01 paratope (Fig. 2). The remaining attention heads scattered attention throughout the structure (Appendix A.7), consistent with previous observations that substantial framework mutations are necessary for functional VRC01 antibodies [10].

## 4 Conclusion

In this work, we explored the use of language models to study the affinity maturation process. Towards this goal, we developed AntiBERTy, an antibody-specific language model trained on a massive set of natural antibody sequences. Next, we showed that immune repertoire sequences encoded with AntiBERTy cluster to resemble affinity maturation trajectories. Closer inspection of these trajectories may provide new biological insights into the affinity maturation process. Finally, we trained a MIL model to detect highly redundant sequences, and showed that the model’s attention localized to binding residues for an extensively studied class of antibodies. In the future, similar methods may enable identification of paratope residues from immune repertoire sequences alone.

## **Acknowledgments and Disclosure of Funding**

The authors thank Jacopo Teneggi, Zhenzhen Wang, Richard Shuai, Dr. Sai Pooja Mahajan, and Dr. Rahel Frick for helpful discussions and advice. This work was supported by National Institutes of Health grants R01-GM078221 and T32-GM008403 (J.A.R.), CISCO research grant CG# 2686384, and AstraZeneca (J.A.R.). Computational resources were provided by the Maryland Advanced Research Computing Cluster (MARCC).

## References

- [1] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M Church, Peter Karl Sorger, and Mohammed N AlQuraishi. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*, 2021.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [3] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.
- [4] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life’s code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.
- [5] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, 32(2):158–168, 2014.
- [6] Brian L Hie, Kevin K Yang, and Peter S Kim. Evolutionary velocity with protein language models. *bioRxiv*, 2021.
- [7] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
- [8] M Gordon Joyce, Masaru Kanekiyo, Ling Xu, Christian Biertümpfel, Jeffrey C Boyington, Stephanie Moquin, Wei Shi, Xueling Wu, Yongping Yang, Zhi-Yong Yang, et al. Outer domain of hiv-1 gp120: antigenic optimization, structural malleability, and crystal structure with antibody vrc-gp04. *Journal of virology*, 87(4):2294–2306, 2013.
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [10] Florian Klein, Ron Diskin, Johannes F Scheid, Christian Gaebler, Hugo Mouquet, Ivelin S Georgiev, Marie Pancera, Tongqing Zhou, Reha-Baris Incesu, Brooks Zhongzheng Fu, et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent hiv-1 neutralization. *Cell*, 153(1):126–138, 2013.
- [11] Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M Deane, and Konrad Krawczyk. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.
- [12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [13] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.
- [14] Daniel Neumeier, Alexander Yermanos, Andreas Agrafiotis, Lucia Csepregi, Tasnia Chowdhury, Roy A Ehling, Raphael Kuhn, Raphaël Brisset-Di Roberto, Mariangela Di Tacchio, Renan Antonialli, et al. Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *bioRxiv*, 2021.
- [15] Sai T Reddy, Xin Ge, Aleksandr E Miklos, Randall A Hughes, Seung Hyun Kang, Kam Hon Hoi, Constantine Chrysostomou, Scott P Hunicke-Smith, Brent L Iverson, Philip W Tucker, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature biotechnology*, 28(9):965–969, 2010.

- [16] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, 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), 2021.
- [17] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):1–8, 2018.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [20] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling Xu, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to hiv-1. *Science*, 329(5993):856–861, 2010.
- [21] Tongqing Zhou, Rebecca M Lynch, Lei Chen, Priyamvada Acharya, Xueling Wu, Nicole A Doria-Rose, M Gordon Joyce, Daniel Lingwood, Cinque Soto, Robert T Bailer, et al. Structural repertoire of hiv-1-neutralizing antibodies targeting the cd4 supersite in 14 donors. *Cell*, 161(6):1280–1292, 2015.
- [22] Tongqing Zhou, Jiang Zhu, Xueling Wu, Stephanie Moquin, Baoshan Zhang, Priyamvada Acharya, Ivelin S Georgiev, Han R Altae-Tran, Gwo-Yu Chuang, M Gordon Joyce, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for hiv-1 neutralization by vrc01-class antibodies. *Immunity*, 39(2):245–258, 2013.

## A Appendix

### A.1 Antibody sequence dataset

The Observed Antibody Space is a database containing over 1B antibody variable domain sequences from 80 immune repertoire sequencing studies. We clustered this set at 95% sequence identity with LinClust [17] to extract a non-redundant set of 588M sequences. Our dataset includes heavy and light chains from six species (human, mouse, rat, camel, rabbit, and rhesus). From the 588M sequences extracted from the OAS database, we hold out 5% for future testing. Of the remaining 95%, we train the model 558M sequences and use 1M for evaluation and hyperparameter tuning.

### A.2 AntiBERTy model

AntiBERTy is based on the BERT transformer encoder model [2] implementation from Huggingface [19]. We set the hidden dimension to 512 and the feedforward dimension to 2048. Our model contains 8 layers, with 8 attention heads per layer. In total, AntiBERTy contains approximately 26M trainable parameters. We train the model for 8 epochs over the full dataset, which takes approximately 10 days when parallelized across four NVIDIA A100 GPUs. Training and evaluation loss are shown in Figure 4.

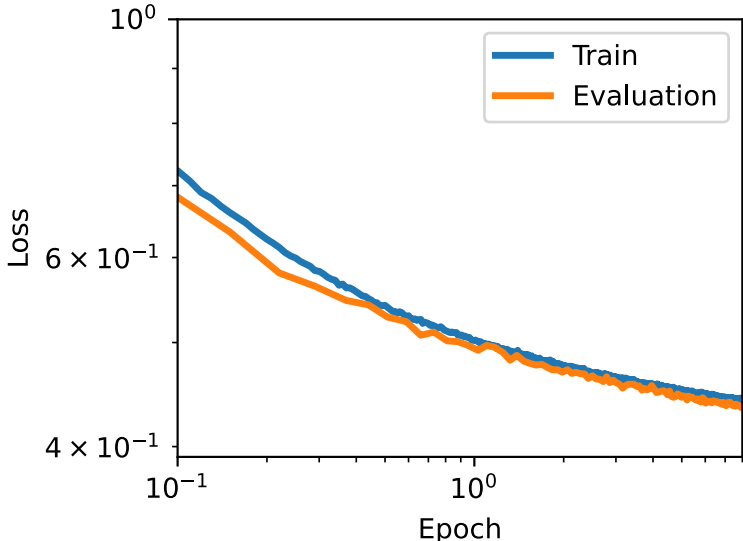


Figure 4: Masked language modeling training and eval loss.

### A.3 HIV-1 donor repertoires

For this work, we collected four previously published immune repertoire sequence datasets from donors who developed neutralizing antibodies against the HIV-1 envelope glycoprotein120. We refer to the four donors using the original identifiers from their respective studies: RU3 [22], IAVI57 [22], IAVI74 [22], and C38 [21]. For each donor, unsorted B-cells were collected from peripheral blood mononuclear cells (PMBC). All sequences are of the IGHG isotype. The number of sequences in each repertoire are given below.



Donor	Number of sequences
RU3	77,067
IAVI57	67,339
IAVI74	40,517
C38	47,670

Table 1: Size of immune repertoire samples for each donor.

#### A.4 Evo-velocity analysis

For each repertoire, we computed evo-velocity scores as described by Hie et al. [6]. For each edge in the KNN graph between two sequences  $x^a$  and  $x^b$ , we calculate an evo-velocity score:

$$v_{ab} = \frac{1}{M} \sum_{i \in M} [\log p(x_i^b | \mathbf{z}_i^a) - \log p(x_i^a | \mathbf{z}_i^b)] \quad (4)$$

where  $M = \{i : x_i^a \neq x_i^b\}$  is the set of residues that differ between  $x^a$  and  $x^b$ , and  $\mathbf{z}_i$  is the latent representation from AntiBERTy when the  $i^{\text{th}}$  residue is masked. In Figure 5, we plot the evo-velocity scores over the Levenshtein-distance-annotated UMAP embeddings for each repertoire. We observe that the evo-velocity arrows consistently align towards the germline sequence rather than in the direction of affinity maturation.

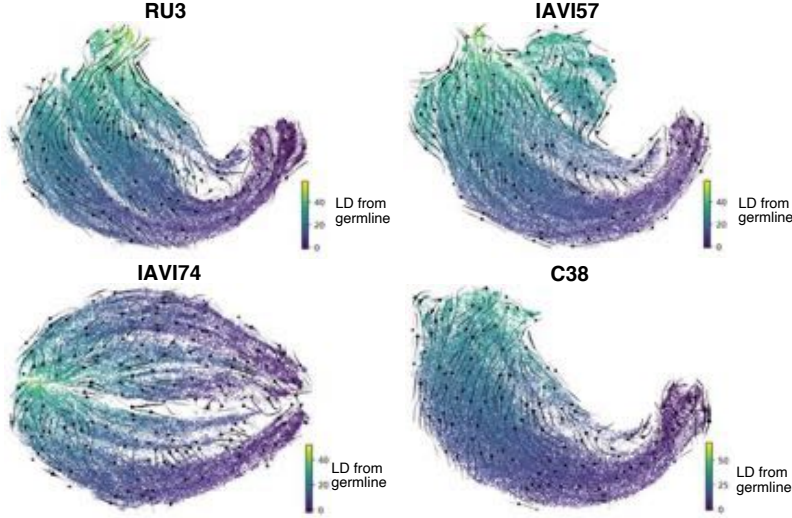


Figure 5: Evo-velocity scores for each repertoire.

#### A.5 MIL model parameters

In Table 2, we report the parameters for the MIL models used in this work. In total, each MIL model contains 31.3K trainable parameters. Models for individual repertoires are trained with batch size 32 and the combined model is trained with batch size 128. All models are trained using the Adam optimizer [9]. The learning rate begins at  $2e-5$  and is decreased according to a cosine annealing schedule.

Parameter	Value
$d_{\text{emb}}$	32
$n_{\text{heads}}$	4
$d_{\text{head}}$	16
$d_{\text{attn}}$	32

Table 2: MIL model parameters

### A.6 MIL model performance metrics

Separate MIL models were trained on sequences from each donor repertoire, as well as a combined dataset of all sequences. During training, 20% of sequences were held out for hyperparameter evaluation. We report performance metrics on this set below.

Donor	Accuracy	AUROC	Precision	Recall	F1 Score
RU3	0.71	0.79	0.83	0.62	0.71
IAVI57	0.90	0.95	0.93	0.89	0.91
IAVI74	0.90	0.96	0.95	0.88	0.91
C38	0.78	0.88	0.81	0.81	0.81
Combined	0.75	0.84	0.74	0.81	0.77

Table 3: Performance metrics for MIL models trained on different repertoire datasets.

### A.7 Other MIL model attention heads

The MIL model trained for this work used four attention heads for embedding variable-length sequences to fixed-sized vectors. Analysis of the second head is provided in the main text (Figure 3). For the remaining heads, we observe attention scattered throughout the antibody (Figure 6). This finding is consistent with observations from previous work that showed substantial framework mutations are necessary to generate functional VRC01 antibodies [10].

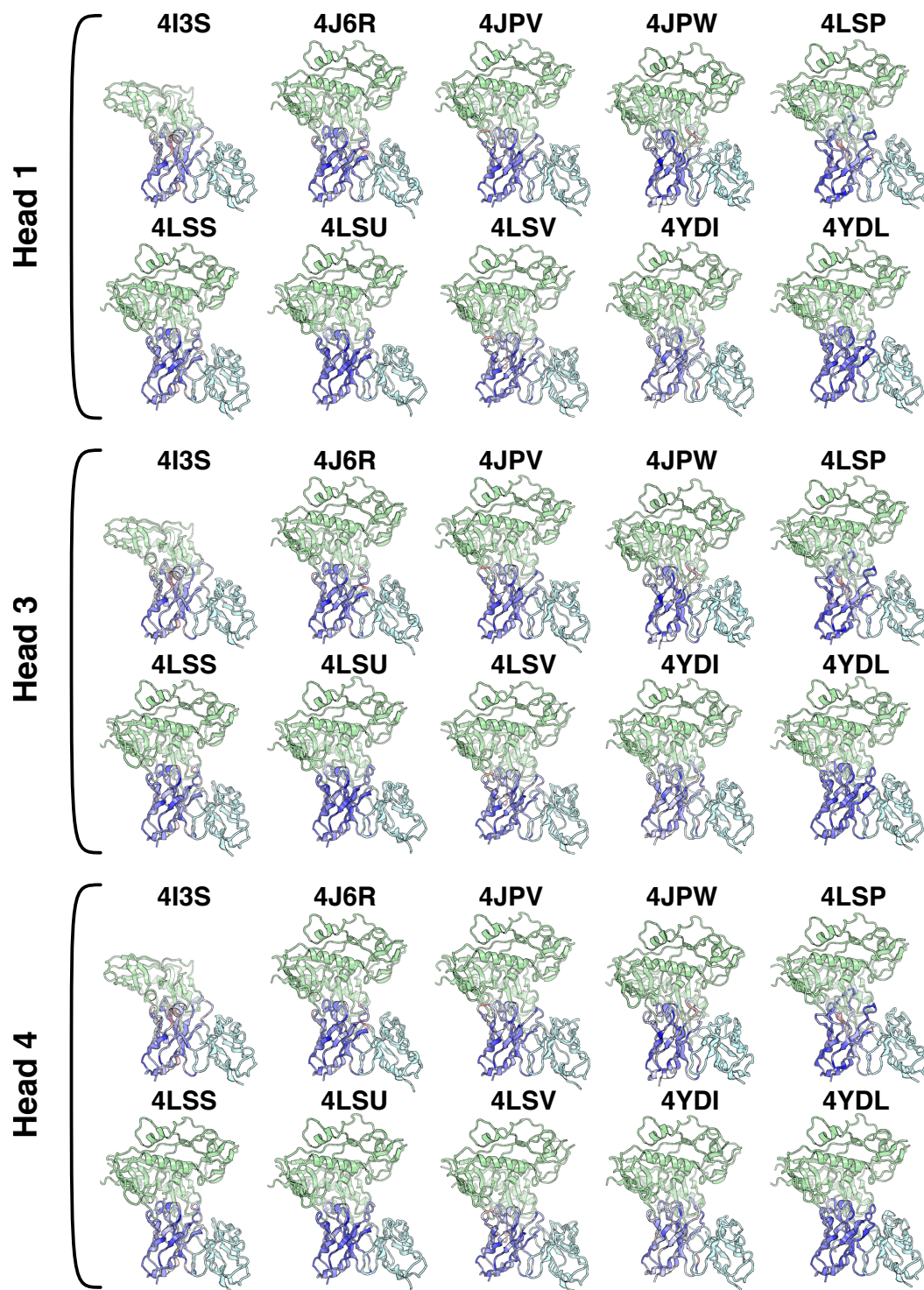


Figure 6: VRC01 antibody complexes annotated with attention from remaining heads. For each structure, the antibody heavy chain is annotated with residue-level attention from MIL model (increasing from blue to red). The light chain and gp120 antigen are shown in cyan and green, respectively.