# Large-scale sequence modeling of antibody-antigen binding specificity

**Anonymous Authors**[1]

## Abstract

Antibody-antigen binding specificity underlies immune protection, vaccine efficacy, and therapeutic antibody development, yet remains difficult to predict. Existing approaches either rely on structure prediction methods ill-suited to antibody-antigen complexes, or sequence-based protein-protein interaction models that require inter-protein coevolution signals absent in antibody-antigen systems. To address this, we develop Agate, a discriminative protein language model trained on over one million human antibody-antigen pairs curated from public datasets, the largest dataset to date. Agate jointly encodes antibody-antigen pairs and is trained using both experimentally validated binders and biologically grounded negative, non-binder examples. To rigorously assess generalization, we implement stringent sequence-based train-test partitioning, including evaluation on completely unseen antigens. Agate achieves state-of-the-art performance for antibody-antigen binding prediction while substantially improving specificity, scalability, and prospective generalization relative to existing methods. Together, these results establish a scalable framework for modeling of antibody recognition with applications in pandemic preparedness, vaccine design, and precision immunotherapy.

## 1. Introduction

The adaptive immune system protects against diverse pathogens through a vast repertoire of antibodies generated through DNA recombination of germline immunoglobulin variable (V), diversity (D), and joining (J) gene segments. Antigen-specific antibody responses result from the activation of naïve B cells with germline-encoded antibodies that recognize an antigen epitope via their six complementarity determining regions (CDRs), followed by affinity maturation and somatic hypermutation to refine binding. Differences in germline immunoglobulin genes across individuals contribute to variation in immune recognition, susceptibility to infectious diseases, and vaccine responsiveness (Collins et al., 2020a; Yuan et al., 2023; deCamp et al., 2024; Sangesland et al., 2022). Convergent antibody responses against shared "public" epitopes further suggest that antibody recognition is constrained by underlying sequence and structural features encoded within the germline repertoire (Shrock et al., 2023).

Despite the central role of germline antibodies in shaping immune responses, predicting antibody-antigen recognition remains a major challenge. Experimental approaches for characterizing germline-encoded antibody responses are difficult to scale and are labor and resource-intensive, requiring B cell isolation, repertoire sequencing, or antigen-specific screening (Shrock et al., 2023). Meanwhile, genetic associations approaches such as Genome-Wide Association Studies (GWAS) have been limited by the complexity of the immunoglobulin loci (Collins et al., 2020b). As a result, the determinants governing which epitopes are intrinsically recognizable by human antibody repertoires remain poorly understood.

While recent advancements in computational approaches have transformed structural biology, accurate antibody-antigen binding specificity remains an unsolved problem (Hitawala & Gray, 2024; Li et al., 2026). Structure-based methods such as AlphaFold for protein-protein interaction prediction (Jumper et al., 2021) face a number of limitations, including: inter-protein coevolution that is largely absent in antibody-antigen binding, limited multiple sequence alignment (MSA) utility for antibodies, hyper-flexible CDR loops, binding through difficult-to-model polar or water-mediated interactions, and residue usage patterns distinct from other heterodimers (McCoy et al., 2024; Li et al., 2026). Alternatively, approaches for antibody design (e.g., IgBert) often encode only the antibody (often even unpaired antibody chains) without antigen-specific information (Kenlay et al., 2024; Olsen et al., 2022b; 2024b; Hepler et al., 2025). As result the dominant approaches in the field are oriented toward designing or ranking likely binders rather than identifying binding specificity (Evans et al., 2021; Tadiello et al., 2026), are generative rather than discriminative (Mille-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

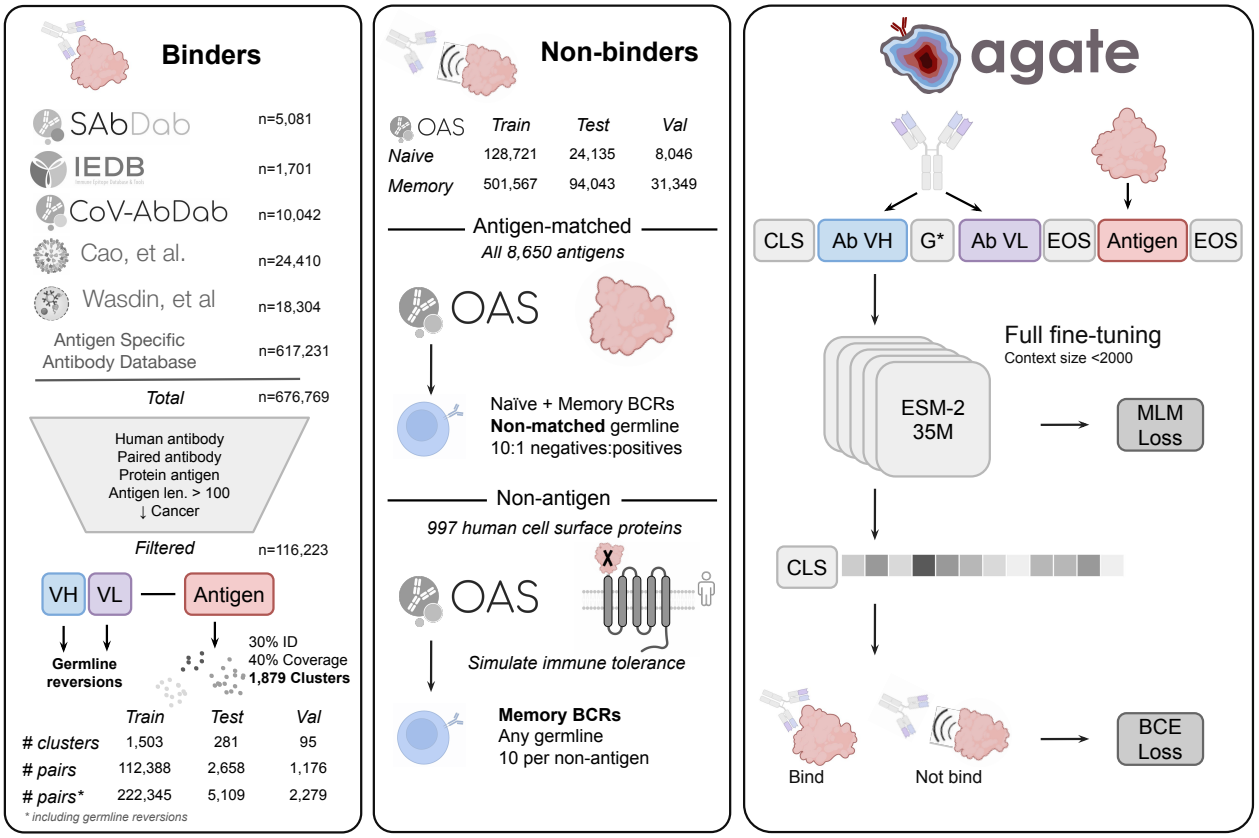


*Figure 1.* **Overview of Agate.** Dataset curation, negative generation, train/test split and model architecture. Ab: Antibody; BCR: B cell receptor; CLS: classification token; G*: Glycine linker; EOS: End of sentence token; MLM: Masked language modeling; BCE: Binary cross entropy; len.: length.

Fragoso et al., 2025; Wasdin et al., 2025; Bennett et al., 2026), and have largely focused on only the most well-studied antigens (Wang et al., 2024; 2022; Bennett et al., 2026). Moreover, structure- or diffusion-based approaches remains computationally impractical for repertoire-scale screening (often $> 10^6$ sequences) or therapeutic library evaluation (Sang et al., 2026).

Sequence-based protein language models (PLMs) provide a scalable alternative and have demonstrated strong performance for general protein-protein interaction prediction (Green et al., 2021; Akiyama et al., 2025; Liu et al., 2024; Ullanat et al., 2025). Yet existing antibody-antigen PLMs frequently encode antibodies and antigens independently using frozen embeddings which limits their ability to learn interaction-specific compatibility features. In addition, many current benchmarks contain substantial sequence overlap between training and evaluation sets, complicating assessment of generalization to unseen antigens (Liu et al., 2024).

To address these limitations, we developed Agate, a discriminative protein language model for antibody-antigen binding prediction trained on over two million human antibody-antigen pairs (Fig. 1). Unlike prior approaches, Agate is jointly trained on antibody-antigen sequence pairs, enabling the model to learn interaction-specific features governing molecular compatibility rather than relying on independently derived representations. Our curated training dataset is the largest collected paired human antibody-antigen binding dataset, with ∼200k more antigen-specific antibodies than structure-based models use from the PDB and several times more positives than used when training other sequence-based antibody-antigen PPI models like MAGE and AntiBinder (Wasdin et al., 2025; Zhang et al., 2024a). To improve binding specificity prediction, we also include biologically grounded negative, non-binder examples derived from B-cell repertoires. We further implement stringent train-test partitioning, including evaluation on completely unseen antigens, to assess model generalization under realistic settings relevant to emerging pathogens and pandemic preparedness.

We show that Agate achieves state-of-the-art performance for antibody-antigen binding prediction while substantially

improving specificity and scalability relative to existing approaches. By learning transferable features governing antibody recognition, Agate enables prediction of epitopes intrinsically recognizable by germline-encoded antibodies and provides a framework for prospective modeling of immune recognition at repertoire scale. These capabilities establish a foundation for applications in vaccine design, therapeutic antibody discovery, immune surveillance, and pandemic preparedness.

## 2. Methods

### 2.1. Dataset curation of antigen-specific antibodies

To enable scalable antibody-antigen specificity modeling, we assembled a large paired antibody-antigen sequence dataset from multiple public resources, including SAbDab, IEDB, and the Antibody Specificity Antigen Database (ASD) (Fig. 1) (Dunbar et al., 2014; Vita et al., 2025; Raybould et al., 2021; Wang et al., 2024; Czerwiński et al., 2026; Wasdin et al., 2025). Compared to structure-based approaches restricted to experimentally resolved complexes, sequence-based modeling permits training on orders of magnitude more antibody-antigen interactions spanning substantially broader antigen diversity.

We restricted the dataset to paired human antibodies containing both the heavy- and light-chain variable domains (VH/VL), excluding non-human antibodies, unpaired chains, and non-protein antigens. In addition to avoid overrepresentation of a small number of highly sampled antigens (e.g., human HER2-associated antibodies represented ∼500,000 pairs in ASD), these were downsampled in the dataset. While the majority of antibodies originated from are human repertoires, the dataset also includes some therapeutically engineered antibodies. Antibodies germline genes were annotated using ANARCI (Dunbar & Deane, 2016). To maximize antigen coverage while remaining compatible with transformer-based architectures, we allowed combined antibody-antigen sequence lengths up to 2,000 residues, exceeding the limits used in previous protein language model interaction frameworks (PLM-interact which uses 1603 and ESM-2 which uses 1024).

Despite integrating multiple datasets, antigen representation remained substantially skewed toward a limited number of intensively studied pathogens and therapeutic targets (e.g., SARS-CoV-2 Spike glycoprotein, Influenza hemagglutinin, and the human cancer target VEGF), reflecting biases inherent to publicly available antibody datasets. To mitigate overfitting and improve generalization, all antigens were clustered at 30% sequence identity and 40% sequence coverage using MMseqs easy-search, resulting in 1,879 clusters (Steinegger & Söding, 2017). Antigen clusters were then divided into training (80%, 1,503 clusters), validation (5%, 95 clusters), and test (15%, 281 clusters) splits to prevent homologous antigens from appearing across partitions and to reduce sequence leakage that can inflate performance estimates in protein interaction prediction benchmarks (Liu et al., 2024; Neumann et al., 2022; Hummer et al., 2025). Moreover, despite significant diversity, there also remains bias in antibody germlines known to impact antibody language model training (Olsen et al., 2024b).

### 2.2. Curation of negative binding set

A further challenge in antibody-antigen specificity prediction is the construction of biologically meaningful negative, non-binder examples. Rather than generating negatives through random shuffling of non-cognate antibody-antigen pairs, which can introduce unrealistic or trivial examples, we generated negatives using two complementary strategies designed to better approximate biologically plausible non-binders.

First, for each antigen in the positive set, we obtain a negative example by sampling naïve and memory BCRs from the Observed Antibody Space (OAS) database (Olsen et al., 2022a) whose germline genes differ from all antibodies known to bind that antigen. Since germline V gene usage is predictive of antigen-binding specificity, given structural motifs that drive recurrent recognition of particular epitopes, differing germlines are less likely to share binding partners (Shrock et al., 2023). We sample 10 negative examples for every positive binding pair.

The second negative set was derived by pairing human memory B cell receptors (from healthy donors in the OAS) with high-confidence human surface proteins (from the Cell Surface Protein Atlas) (Bausch-Fluck et al., 2015). These pairs were treated as putative non-binders based on immune tolerance mechanisms that eliminate or suppress strongly self-reactive B cells during development. Consequently, B cell receptors recognizing endogenous human surface proteins are unlikely to persist in healthy donor repertoires (Nemazee, 2017).

Before pairing with different antigens for negatives, OAS antibodies were split into train (80%), validation (5%), or test (15%) sets to prevent the same antibody from appearing as negatives in multiple sets. In total, we train on 222,345 positive and 1,969,807 negative antibody-antigen pairs, across 1,503 antigen clusters. Agate's positive set is 12 times larger than for MAGE and almost four times larger for AntiBinder.

Information on training, validation and test splits including VH/VL pairings, V gene germline identity of heavy and light chains, and antigen taxonomies are summarized in Figure S1.

### 2.3. Model framework and training

Agate jointly encodes antibody-antigen sequence pairs by concatenating their sequences as input to the encoder-only PLM, ESM-2, which we fine-tune end-to-end (Fig 1). Following PLM-interact (Liu et al., 2024), the training objective combines continued masked language modelling with a binary classification loss in a 1:10 ratio. The original ESM tokenizer jointly encodes sequences concatenated as [CLS][Antibody VH][GGGGG][Antibody VL][EOS][Antigen][EOS], using a flexible glycine linker to separate the antibody chains rather than an additional EOS token. Classification is applied via a linear head on the classification [CLS] ESM-2 output embeddings, predicting whether each pair interacts. To prevent bias from antigen overrepresentation, we use a custom batch sampler that samples each antigen cluster with uniform probability and groups sequences by length to reduce padding. The model was trained for 18 epochs (where one epoch is 50 samples per antigen) on 1 L40S GPU over approximately 42 hours, using the 35M FAESM (Fred Zhangzhi Peng & contributors, 2024) for efficiency.

### 2.4. Training on Spike-excluded set and epitope determination

To determine how Agate would perform at predicting binders for the SARS-CoV-2 Spike glycoprotein, a version of Agate was trained completely excluding any Spike glycoprotein sequence from any strain of severe acute respiratory syndrome coronavirus. In total, 40,888 antibody-antigen pairs were excluded, resulting in 99,853 positive sequences for training and 2,163 in validation (Fig 3). Including negatives in the same way outlined above, there were a total of 1,679,673 training sequences. The model was trained for 18 epochs on 1 GPU for approximately 42 hours.

In addition, we wanted to see if Agate had some signal for predicting binding contacts from sequence information alone. To do so, we computed per-residue attention scores from the CLS token to antigen residues. Attention was averaged across the final transformer layers and then projected onto a canonical SARS-CoV-2 receptor binding domain (RBD) reference sequence via local alignment. Cluster-level consensus signal was computed by taking the mean attention across all antigen sequences within the same cluster. The final per-complex score combined the per-complex attention signal with this cluster-level consensus signal. Ground-truth epitopes were defined from antibody-antigen heavy-atom contacts in PDB structures. Epitope prediction performance was evaluated using per-complex AUROC against structure-derived epitope labels and compared against ESM2 (35M), Agate trained without Spike, and Agate trained with Spike using splits outlined above.

## 3. Results

We benchmarked Agate against four existing protein interaction models: PLM-Interact (Liu et al., 2024), MINT (Ullanat et al., 2025), MAGE (Wasdin et al., 2025), and Boltz-2 (Passaro et al., 2025). Similar to Agate, MINT, PLM-Interact, and MAGE are fine-tuned protein language models (PLMs). MINT and PLM-Interact both extend ESM-2 (Lin et al., 2023), whereas MAGE fine-tunes ProGen2 (Nijkamp et al., 2023). MINT incorporates cross-chain attention and is trained on approximately 96 million experimentally derived and predicted protein-protein interactions from STRING-DB. PLM-Interact jointly encodes protein pairs and is fine-tuned on multiple interaction datasets, including virus-host interactions from the Host-Pathogen Interaction Database and human protein interactions from STRING-DB, using both cognate binding pairs and shuffled non-cognate negatives. In contrast, MAGE is specifically trained for antibody-antigen prediction using only positive examples, with 18,507 curated paired antibody-antigen sequences. We additionally evaluated Boltz-2, an AlphaFold-like structure-based model trained using PDB structures, molecular dynamics ensembles, and approximately five million affinity measurements.

Binding scores for MINT were calculated using the provided code and downstream model checkpoint for Binary PPI classification on their GitHub repository. Because MAGE is a generative model, we estimate the likelihood of an antibody conditioned on a query antigen by computing the perplexity over antibody tokens, given the antigen as a prefix. For Boltz-2, the ipTM score was used as the binding score.

### 3.1. Agate outperforms existing antibody-antigen prediction models

We first evaluated each model on its ability to distinguish cognate antibody-antigen interactions from non-binding pairs in the held-out test set (n = 72,479). Agate substantially outperformed all baseline models, achieving an AUC of 0.97, compared to 0.58 for Boltz-2, 0.54 for both MAGE and PLM-Interact, and 0.45 for MINT (Fig. 2A). The performance advantage of Agate persisted when evaluated using precision-recall analysis, indicating robust discrimination despite substantial class imbalance within the evaluation dataset (Fig. 2B).

To determine whether Agate's improved performance was driven by a small number of overrepresented antigens, we next evaluated model performance independently across antigen groups. For each antigen, we quantified the separation between positive and negative interaction scores using the point-biserial correlation coefficient. Agate consistently outperformed all comparator models across nearly all antigens in the test set, demonstrating that performance gains were broadly distributed rather than dominated by a subset

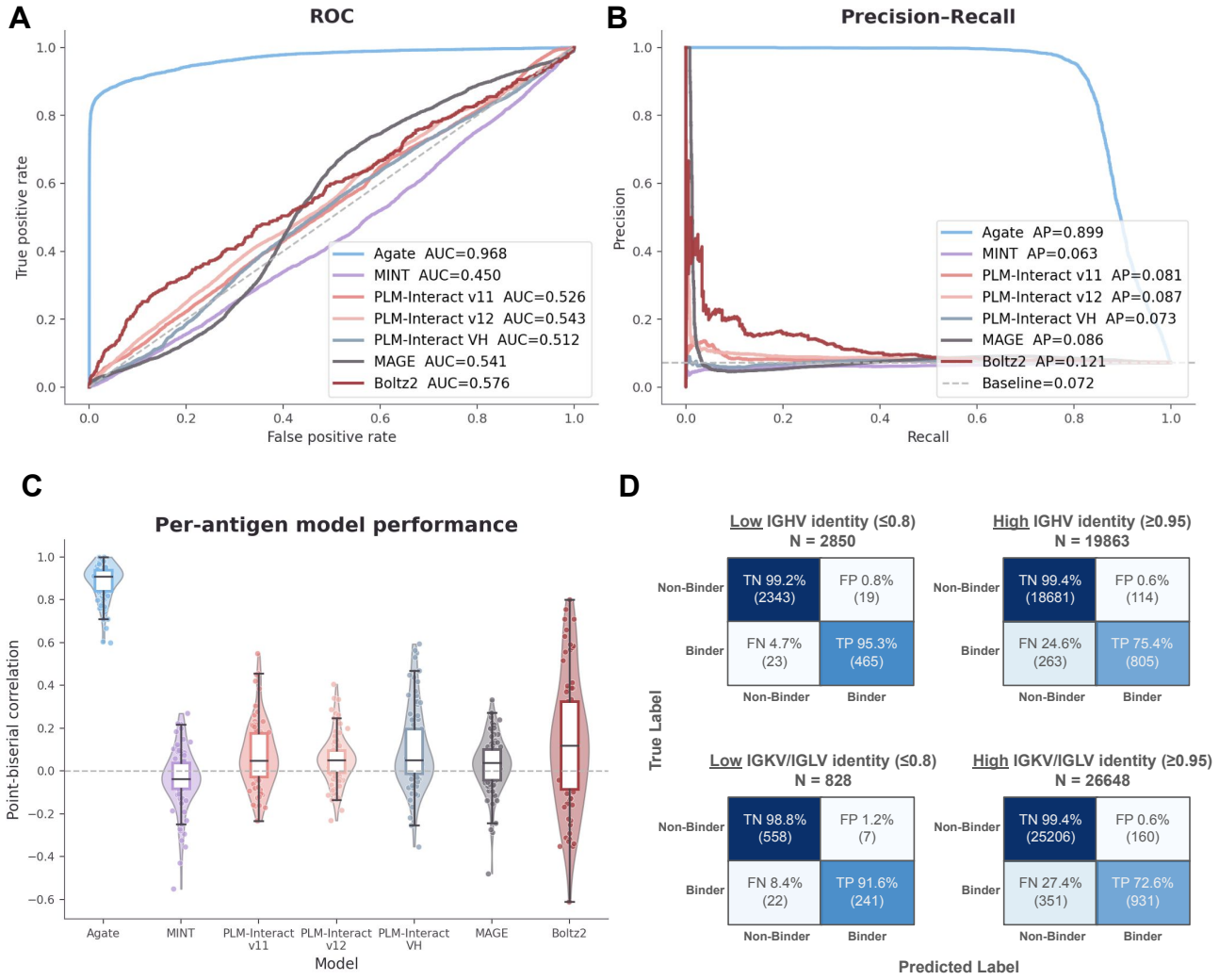


*Figure 2.* **Agate performance summary.** **(A)** ROC relative to the benchmark models MINT, PLM-Interact, MAGE, and Boltz-2. PLM-interact VH, V11, V12 represent the three model checkpoints trained on human PPIs, virus-human PPIs, and leakage-free human PPIs, respectively (Liu et al., 2024). **(B)** PR relative to the benchmark models MINT, PLM-Interact, MAGE, and Boltz-2. The baseline classifier represents the fraction of positives. **(C)** Per-antigen model performance, filtered on any unique antigen with more than ten known antibody binders, compared to benchmark models. **(D)** Confusion matrix subset to assess Agate performance on antibodies that have a low and high germline identity.

of highly represented targets (Fig. 2C).

Because germline-encoded antibody features play an important role in antigen recognition, we next examined model performance as a function of antibody divergence from germline sequence (Fig 2D). Agate maintained strong performance across all levels of somatic hypermutation, but achieved its highest sensitivity for antibodies with lower germline similarity. For antibodies with $\leq 80\%$ sequence identity to germline, Agate achieved a true positive rate of 95.3 and 91.6 % for heavy versus light V genes respectively, while maintaining a true negative rate above 99.2 and 98.8%. In contrast, performance decreased for highly germline-like antibodies ($\geq 95\%$ germline identity), where the true positive rate declined to 75.4 (heavy) and 72.6% (light) despite maintaining high specificity (true negative rate $> 99.4\%$).

This reduction in sensitivity for germline-like antibodies likely reflects properties of the negative sampling strategy used during training, where naïve B-cell receptors were incorporated as putative non-binders with a ration of 10 non-binders for every one true binder. This strategy or high ratio of negatives may bias the model toward classifying highly germline-proximal antibodies as non-binding, and may artificially inflate the performance of Agate relative to benchmarked methods not trained with this assumption. Future work incorporating experimentally validated naïve antibody interactions or alternative negative genera-

tion strategies may further improve sensitivity in this regime Nevertheless, Agate substantially outperformed all existing approaches across both germline-diverged and germline-like antibodies.

### 3.2. Agate identifies SARS-CoV-2 Spike binding contacts

We next investigated whether Agate could identify not only whether an antibody and antigen interact, but also which residues are likely to mediate binding. We focused this analysis on the SARS-CoV-2 Spike protein because of the large number of experimentally resolved antibody-Spike complexes available for evaluation.

To rigorously assess model generalization and avoid data leakage, we trained a new version of Agate using the same architecture and training procedure as the primary model, but with all coronavirus Spike glycoprotein sequences removed from the training set (Fig 3A). Despite never observing Spike during training, this Spike-excluded Agate model achieved strong binding prediction performance on held-out Spike antibody-antigen pairs (AUROC = 0.79), outperforming MAGE (AUROC = 0.71), which was trained on a dataset with a majority ofantibodies binding to SARS-CoV-related proteins (Fig 3B) (Wasdin et al., 2025).

Drawing inspiration from previous work demonstrating that PLMs can code for residue-residue dependencies correlating to structural contacts, we asked whether Agate learned residue-level interaction features associated with physical antibody-antigen contacts (Zhang et al., 2024b). Because Agate was trained to identify binders using the CLS token, we hypothesized that CLS attention magnitudes would reflect key sequence residues for epitope binding. To identify putative binding residues, we computed per-residue attention scores from the CLS token to antigen residues and quantified whether they corresponded to experimentally observed interface contacts from solved structures. Spike-excluded Agate substantially outperformed the base ESM-2 model in recovering true Spike residues contacting antibodies (AUC = 0.743 versus 0.674), indicating that fine-tuning on antibody-antigen interaction data enables the model to learn interaction-specific structural compatibility signals beyond those captured by general protein language modeling alone (Fig 3D-F).

Including Spike glycoprotein sequences during Agate training further improved contact prediction performance (AUC = 0.784), suggesting that the model benefits both from transferable interaction priors learned across antibody-antigen systems and from antigen-specific examples when available. Together, these results demonstrate that Agate captures biologically meaningful residue-level interaction information despite being trained only on sequence-level binding supervision.

## 4. Conclusions and future directions

In this work, we aimed to develop a model of antibody-antigen binding capable of prediction at repertoire scale. To achieve this, we trained a PLM that jointly encodes antibody-antigen pairs on both experimentally validated binders biologically grounded negative examples. Agate achieves state-of-the-art performance for antibody-antigen prediction across several benchmarks and preliminary results indicate CLS attention magnitudes correspond to epitope binding residues.

There are several directions we plan to pursue for future work, one of which is further developing the model framework. To make training more efficient we plan to sample CDR residues in the antibody sequence more frequently for masking. We also plan to explore different weightings for both the MLM and classification loss during the training. In addition, we are interested in including structure information as additional features to the model. Another avenue of future work is enhancing the datasets we use for training and validation. In particular, we are interested in using stronger sources of negatives, such as experimentally determined negative binders in LIBRA-seq datasets (Abu-Shmais et al., 2024). We are also interested in benchmarking our model with traditional methods of antibody negatives via shuffling of antibodies positives for alternate antigens, to avoid inflating our own model performance by training and validating on the same negative distribution. In addition, we would like to interrogate the germline bias present in our model, and incorporate appropriate changes as necessary (Olsen et al., 2024a). At this stage, Agate more often incorrectly assigns antibodies with higher germline identifies as negative, which is likely a function of more frequent germline-like antibodies present as negative examples in the training set (Fig. S1A). Finally, we would like to evaluate several recent and relevant models as baselines — namely Origin-1 and Protenix, which have demonstrated improvements in antibody–antigen structure prediction, particularly through extensive sampling(Levine et al., 2026; Zhang et al., 2026) as well as other metrics like ipSAE — and assess the extent to which these gains transfer to binding specificity prediction.

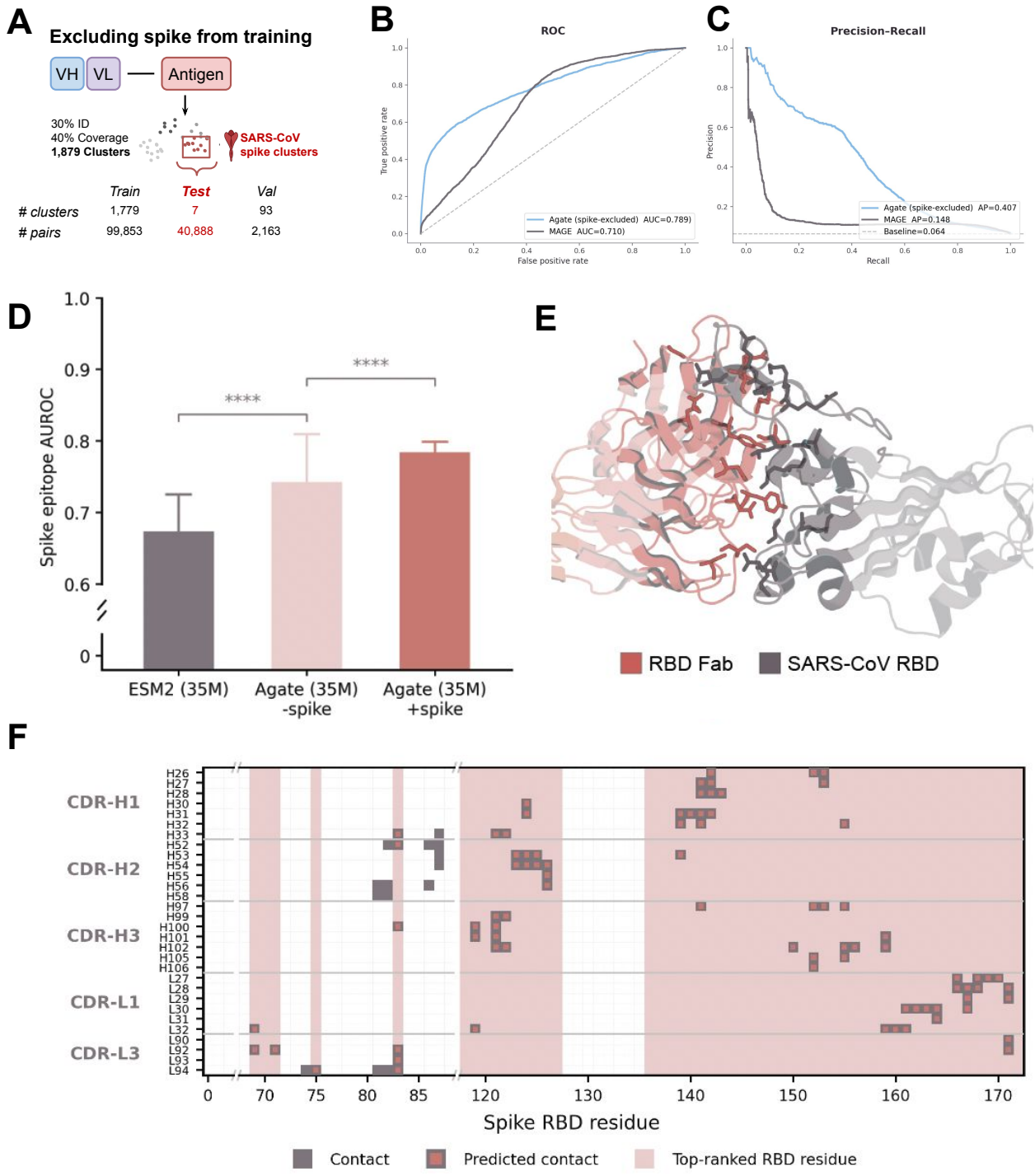


*Figure 3.* **Results from Spike-excluded trained Agate model and RBD epitope determination (A)** Dataset curation and splits so all SARS-CoV Spike glycoprotein antigens and their respective antibodies were exclusively in the test set. **(B)** ROC curve comparing Spike-excluded Agate to MAGE, a generative antigen-specific paired antibody PLM with significant training on CoV-specific antibodies. **(C)** Precision-recall curve comparing Spike-excluded Agate to MAGE. **(D)** Spike epitope AUROC of ESM2 compared to both versions of Agate. **(E)** Representative complex of a human antibody bound to SARS-CoV-2 RBD (RCSB id: 7QF0) and their contacts. **(F)** Residue-level contact map grouped by Spike RBD residue (numbering within RBD alone) and IMGT residues, colored by true contacts, predicted contacts that overlap with true residues, and highlighted top-ranked antigen residues across the sequence.

## References


Abu-Shmais, A. A., Vukovich, M. J., Wasdin, P. T., Suresh, Y. P., Marinov, T. M., Rush, S. A., Gillespie, R. A., Sankhala, R. S., Choe, M., Joyce, M. G., Kanekiyo, M., McLellan, J. S., and Georgiev, I. S. Antibody sequence determinants of viral antigen specificity. *MBio*, 15(10): e0156024, October 2024.

Akiyama, Y., Zhang, Z., Mirdita, M., Steinegger, M., and Ovchinnikov, S. Scaling down protein language modeling with msa pairformer. *bioRxiv*, pp. 2025–08, 2025.

Bausch-Fluck, D., Hofmann, A., Bock, T., Frei, A. P., Cerciello, F., Jacobs, A., Moest, H., Omasits, U., Gundry, R. L., Yoon, C., et al. A mass spectrometric-derived cell surface protein atlas. *PloS one*, 10(4):e0121314, 2015.

Bennett, N. R., Watson, J. L., Ragotte, R. J., Borst, A. J., See, D. L., Weidle, C., Biswas, R., Yu, Y., Shrock, E. L., Ault, R., et al. Atomically accurate de novo design of antibodies with rfdiffusion. *Nature*, 649(8095):183–193, 2026.

Collins, A. M., Yaari, G., Shepherd, A. J., Lees, W., and Watson, C. T. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? *Curr. Opin. Syst. Biol.*, 24:100–108, December 2020a.

Collins, A. M., Yaari, G., Shepherd, A. J., Lees, W., and Watson, C. T. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? *Current opinion in systems biology*, 24:100–108, 2020b.

Czerwiński, A., Dudzic, P., Wójtowicz, K., Jaszczyszyn, I., Bielska, W., Wrobel, S., Demharter, S., Spreafico, R., Greiff, V., and Krawczyk, K. Asd: antigen-specific antibody database. In *Mabs*, volume 18, pp. 2623330. Taylor & Francis, 2026.

deCamp, A. C., Corcoran, M. M., Fulp, W. J., Willis, J. R., Cottrell, C. A., Bader, D. L. V., Kalyuzhniy, O., Leggat, D. J., Cohen, K. W., Hyrien, O., Menis, S., Finak, G., Ballweber-Fleming, L., Srikanth, A., Plyler, J. R., Rahaman, F., Lombardo, A., Philiponis, V., Whaley, R. E., Seese, A., Brand, J., Ruppel, A. M., Hoyland, W., Mahoney, C. R., Cagigi, A., Taylor, A., Brown, D. M., Ambrozak, D. R., Sincomb, T., Mullen, T.-M., Maenza, J., Kolokythas, O., Khati, N., Bethony, J., Roederer, M., Diemert, D., Koup, R. A., Laufer, D. S., McElrath, J. M., McDermott, A. B., Karlsson Hedestam, G. B., and Schief, W. R. Human immunoglobulin gene allelic variation impacts germline-targeting vaccine priming. *NPJ Vaccines*, 9(1):58, March 2024.

Dunbar, J. and Deane, C. M. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

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

Evans, R., O'neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Fred Zhangzhi Peng, P. C. and contributors. Faesm: An efficient pytorch implementation of evolutionary scale modeling (esm). https://github.com/pengzhangzhi/faesm, 2024. Efficient PyTorch implementation of ESM with FlashAttention and Scalar Dot-Product Attention (SDPA).

Green, A. G., Elhabashy, H., Brock, K. P., Maddamsetti, R., Kohlbacher, O., and Marks, D. S. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature communications*, 12(1):1396, 2021.

Hepler, N. L., Hill, A. J., Jaffe, D. B., Gibbons, M. C., Pfeiffer, K. A., Hilton, D. M., Freeman, M., and McDonnell, W. J. Better antibodies engineered with a glimpse of human data. *bioRxiv*, pp. 2025–06, 2025.

Hitawala, F. N. and Gray, J. J. What has AlphaFold3 learned about antibody and nanobody docking, and what remains unsolved? *bioRxiv*, pp. 2024.09.21.614257, September 2024.

Hummer, A. M., Schneider, C., Chinery, L., and Deane, C. M. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\delta\delta$ g prediction. *Nature Computational Science*, 5(8):635–647, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., and Deane, C. M. Large scale paired antibody language models. *PLOS Computational Biology*, 20(12): e1012646, 2024.

Levine, S., King, J. E., Stern, J., Grayson, D., Wang, R., Yin, R., Lupo, U., Kulyte, P., Brand, R. M., Bertin, T., Pfingsten, R., Cejovic, J., Chung, C., Luton, B. K., Hagemann, A., Haile, R., Medina, E., Panwar, P., Dubrovskyi, O., LaCombe, C., Anderson, Z., Mildh, D., Benjamin, S., Kaiser, J., Ferron, J., Sarrico, M., Kershner, A., Mishra, A., Ejan, K. R., Marsh, E. K., Bringas, P., Vilaychack, P., Chapman, K., Ripley, J., Gowda, M., Collins, K. M.,

McCloskey, C. M., Joseph, J. S., Ripley, R., Abdulhaqq, S. A., Spencer, D., Devine, T., Feltner, A., Guerin, M., Goby, J., Hendricks, J., Castillo, D., McClain, S., Ganini, D., Shpiel, D., Mategko, J., Garcia, E. C., Zabet-Moghaddam, M., Sutton, J. M., Guo, Z., West, S. M., Iyer, J. S., and Shanehsazzadeh, A. Origin-1: a generative AI platform for de novo antibody design against novel epitopes. *bioRxiv*, pp. 2026.01.14.699389, January 2026.

Li, S., Mu, Z., and Yan, C. Dissecting the black box of alphafold in protein-protein complex assembly. *bioRxiv*, pp. 2026–04, 2026.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

Liu, D., Young, F., Lamb, K. D., Claudio Quiros, A., Pancheva, A., Miller, C., Macdonald, C., Robertson, D. L., and Yuan, K. PLM-interact: extending protein language models to predict protein-protein interactions. *bioRxiv*, pp. 2024.11.05.622169, November 2024.

McCoy, K. M., Ackerman, M. E., and Grigoryan, G. A comparison of antibody-antigen complex sequence-to-structure prediction methods and their systematic biases. *Protein Sci.*, 33(9):e5127, September 2024.

Mille-Fragoso, L. S., Wang, J. N., Driscoll, C. L., Dai, H., Widatalla, T., Zhang, X., Hie, B. L., and Gao, X. J. Efficient generation of epitope-targeted de novo antibodies with germinal. *bioRxiv*, 2025.

Nemazee, D. Mechanisms of central tolerance for b cells. *Nature Reviews Immunology*, 17(5):281–294, 2017.

Neumann, D., Roy, S., Minhas, F. U. A. A., and Ben-Hur, A. On the choice of negative examples for prediction of host-pathogen protein interactions. *Frontiers in Bioinformatics*, 2:1083292, 2022.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Syst*, 14(11):968–978.e3, November 2023.

Olsen, T. H., Boyles, F., and 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, 2022a.

Olsen, T. H., Moal, I. H., and Deane, C. M. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022b.

Olsen, T. H., Moal, I. H., and Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv*, pp. 2024.02.02.578678, February 2024a.

Olsen, T. H., Moal, I. H., and Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11): btae618, 2024b.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Raybould, M. I., Kovaltsuk, A., Marks, C., and Deane, C. M. Cov-abdab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, 2021.

Sang, Z., Xiang, Y., Huang, W., Sargunas, P. R., Kim, Y. J., Feng, Z., Taylor, D. J., and Shi, Y. Repertoire-scale antibody structural prediction informs therapeutic design. *Science Advances*, 12(17):eaef7163, 2026.

Sangesland, M., Torrents de la Peña, A., Boyoglu-Barnum, S., Ronsard, L., Mohamed, F. A. N., Moreno, T. B., Barnes, R. M., Rohrer, D., Lonberg, N., Ghebremichael, M., Kanekiyo, M., Ward, A., and Lingwood, D. Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly neutralizing antibodies against influenza virus. *Immunity*, 55(9):1693–1709.e8, September 2022.

Shrock, E. L., Timms, R. T., Kula, T., Mena, E. L., West, Jr, A. P., Guo, R., Lee, I.-H., Cohen, A. A., McKay, L. G. A., Bi, C., Leng, Y., Fujimura, E., Horns, F., Li, M., Wesemann, D. R., Griffiths, A., Gewurz, B. E., Bjorkman, P. J., and Elledge, S. J. Germline-encoded amino acid-binding motifs drive immunodominant public antibody responses. *Science*, 380(6640):eadc9498, April 2023.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Tadiello, M., Ludaic, M., Viliuga, V., and Elofsson, A. Abagrank: Improving model selection of alphafold antibody-antigen complexes by learning to rank. *bioRxiv*, pp. 2026–03, 2026.

Ullanat, V., Jing, B., Sledzieski, S., and Berger, B. Learning the language of protein–protein interactions. *bioRxiv*, pp. 2025.03.09.642188, March 2025.

Vita, R., Blazeska, N., Marrama, D., Marcus, I. C. T. M. S. D. Z. L. F. G. Z. L. C. K. R. B. F. S. B. M. C. S. S. C. E. L. M. C. E., Duesing, S., Bennett, J., Greenbaum,

J., De Almeida Mendes, M., Mahita, J., Wheeler, D. K., et al. The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443, 2025.

Wang, Y., Yuan, M., Lv, H., Peng, J., Wilson, I. A., and Wu, N. C. A large-scale systematic survey reveals recurring molecular features of public antibody responses to sars-cov-2. *Immunity*, 55(6):1105–1117, 2022.

Wang, Y., Lv, H., Teo, Q. W., Lei, R., Gopal, A. B., Ouyang, W. O., Yeung, Y.-H., Tan, T. J., Choi, D., Shen, I. R., et al. An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies. *Immunity*, 57(10):2453–2465, 2024.

Wasdin, P. T., Johnson, N. V., Janke, A. K., Held, S., Marinov, T. M., Jordaan, G., Gillespie, R. A., Vandenabeele, L., Pantouli, F., Powers, O. C., et al. Generation of antigen-specific paired-chain antibodies using large language models. *Cell*, 188(25):7206–7221, 2025.

Yuan, M., Feng, Z., Lv, H., So, N., Shen, I. R., Tan, T. J. C., Teo, Q. W., Ouyang, W. O., Talmage, L., Wilson, I. A., and Wu, N. C. Widespread impact of immunoglobulin V-gene allelic polymorphisms on antibody reactivity. *Cell Rep.*, 42(10):113194, October 2023.

Zhang, K., Tao, Y., and Wang, F. AntiBinder: utilizing bidirectional attention and hybrid encoding for precise antibody-antigen interaction prediction. *Brief. Bioinform.*, 26(1):bbaf008, November 2024a.

Zhang, Y., Gong, C., Sun, J., Guan, J., Ren, M., Xue, S., Zhang, H., Ma, W., Liu, Z., Chen, X., and Xiao, W. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717613. URL https://www.biorxiv.org/content/early/2026/04/11/2026.04.10.717613.

Zhang, Z., Wayment-Steele, H. K., Brixi, G., Wang, H., Peraro, M. D., Kern, D., and Ovchinnikov, S. Protein language models learn evolutionary statistics of interacting sequence motifs. *bioRxiv*, pp. 2024.01.30.577970, January 2024b.

# A. Supplemental Figures

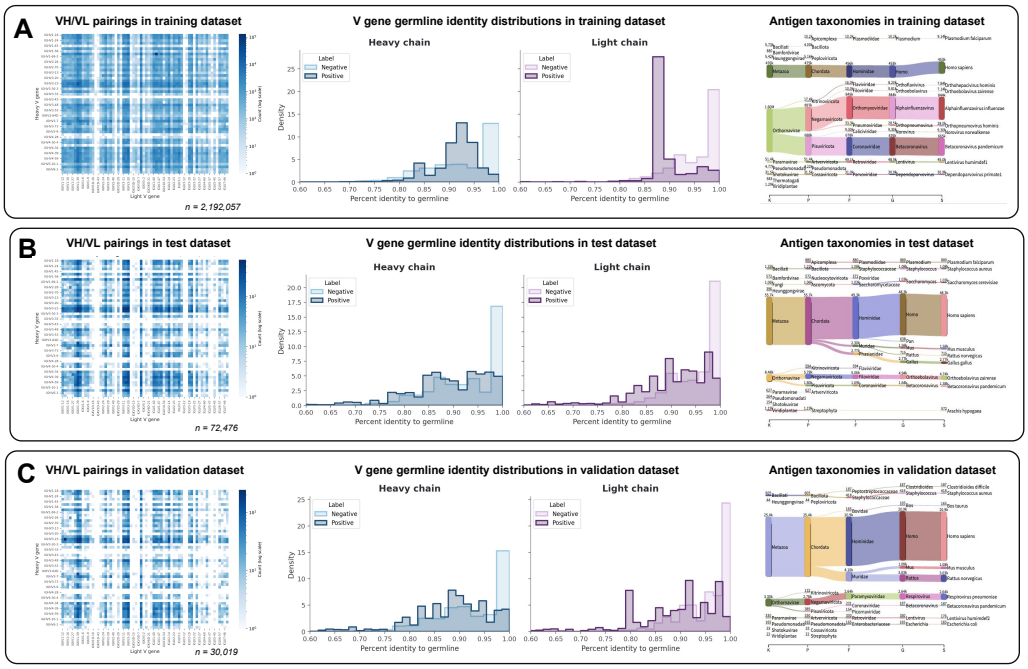


*Figure S1.* **Composition of Agate dataset splits** VH/VL pairing, germline identity relative to closest mapped human V gene from IMGT, and antigen taxonomies for **(A)** train, **(B)** test, and **(C)** validation.