Article

Immunity

An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies

Highlights

- Assembled a dataset of 5,561 published antibodies to influenza HA from 132 donors
- Antibodies to HA head and stem domains have distinct convergent sequence features
- Developed a lightweight language model (mBLM) for antibody specificity prediction
- Discovered HA stem antibodies and key somatic hypermutations using mBLM

Authors

Yiquan Wang, Huibin Lv, Qi Wen Teo, ..., Xin Chen, Claire S. Graham, Nicholas C. Wu

Correspondence

nicwu@illinois.edu

In brief

Predicting antibody specificity based solely on sequence has remained an obstacle in humoral research. Leveraging a curated dataset of >5,000 influenza hemagglutinin (HA) antibodies, Wang et al. develop a lightweight memory B cell language model for antibody specificity prediction. This model identifies unique HA stem antibodies and key antibody sequence features.

Wang et al., 2024, Immunity *57*, 2453–2465
October 8, 2024
https://doi.org/10.1016/j.immuni.2024.07.022

CellPress

Immunity

CellPress

Article

An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies

Yiquan Wang,[1,7] Huibin Lv,[1,2,7] Qi Wen Teo,[1,2] Ruipeng Lei,[1] Akshita B. Gopal,[1] Wenhao O. Ouyang,[1] Yuen-Hei Yeung,[1,3,4] Timothy J.C. Tan,[5] Danbi Choi,[1] Ivana R. Shen,[1] Xin Chen,[5] Claire S. Graham,[1] and Nicholas C. Wu[1,2,5,6,8,*]

[1]Department of Biochemistry, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[2]Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[3]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[4]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[5]Center for Biophysics and Quantitative Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[6]Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: nicwu@illinois.edu
https://doi.org/10.1016/j.immuni.2024.07.022

SUMMARY

Despite decades of antibody research, it remains challenging to predict the specificity of an antibody solely based on its sequence. Two major obstacles are the lack of appropriate models and the inaccessibility of datasets for model training. In this study, we curated >5,000 influenza hemagglutinin (HA) antibodies by mining research publications and patents, which revealed many distinct sequence features between antibodies to HA head and stem domains. We then leveraged this dataset to develop a lightweight memory B cell language model (mBLM) for sequence-based antibody specificity prediction. Model explainability analysis showed that mBLM could identify key sequence features of HA stem antibodies. Additionally, by applying mBLM to HA antibodies with unknown epitopes, we discovered and experimentally validated many HA stem antibodies. Overall, this study not only advances our molecular understanding of the antibody response to the influenza virus but also provides a valuable resource for applying deep learning to antibody research.

INTRODUCTION

Discovery and characterization of monoclonal antibodies are central to the understanding of human immune response, as well as the design of vaccines and therapeutics.[1,2] As exemplified by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) research in the past few years, antibody discovery has dramatically accelerated due to the technological advancements in single-cell high-throughput screen[3] and paired B cell receptor sequencing.[4] Nevertheless, epitope mapping remains a major bottleneck of antibody characterization, which often involves the determination of individual antigen-antibody complex structures using X-ray crystallography or cryoelectron microscopy (cryo-EM). As a result, there is a huge interest in developing methods for antibody specificity prediction.

Despite the huge diversity of human antibody repertoire, with at least 10^{15} antibody sequences,[5,6] antibody responses from different individuals often utilize recurring sequence features to target a given epitope.[7–15] This phenomenon is also known as convergent or public antibody response. Traditionally, antibody specificity prediction has mainly relied on biophysical models.[16] However, the observation of public antibody response suggests that antibody specificity prediction can also be achieved by an orthogonal, data-driven approach. Specifically, with a sufficiently large sequence dataset of human antibodies that share a common epitope, a purely sequence-based model can be trained to predict whether an antibody targets this given epitope or not.

The application of natural language processing has revolutionized protein structure and function prediction as well as protein design.[17–23] Although there are several language models for antibodies,[24–26] to the best of our knowledge, none of them enables antibody specificity prediction. One of the major barriers to developing a language model for antibody specificity prediction is the lack of systematically assembled datasets for model training, which would require both sequence and epitope information for individual antibodies. Although many studies have reported sequences of antibodies with known epitopes, such information is often not centralized. Databases such as CoV-AbDab, which documents the sequence and epitope information for >10,000 antibodies to coronavirus,[27] are absent for most pathogens, including influenza virus.

Hemagglutinin (HA) is the major antigen of influenza virus and has a hypervariable globular head domain atop a highly

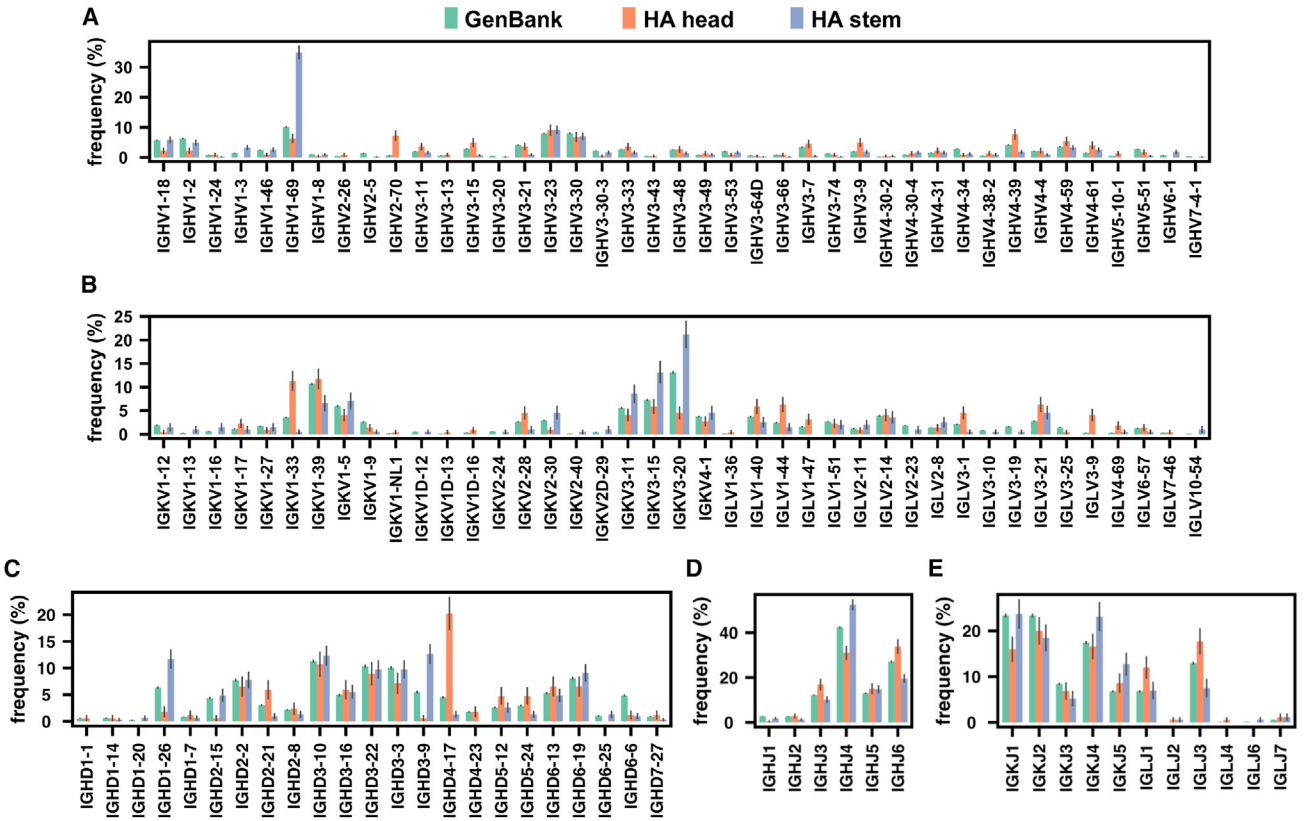


Figure 1. Germline gene usages in influenza HA antibodies

(A) The IGHV gene usage, (B) IGK(L)V gene usage, (C) IGHD gene usage, (D) IGHJ gene usage, and (E) IGK(L)J gene usage in antibodies to HA head domain (orange) and HA stem domain (blue). For comparison, germline gene usages of all antibodies from GenBank are also shown (green). To avoid being confounded by B cell clonal expansion, a single clonotype from the same donor is considered as one antibody (see STAR Methods). Error bars represent the standard deviation computed from binomial distribution.

conserved stem domain.[28] In this study, we manually curated 5,561 human antibodies to influenza HA protein from research publications and patents. Recurring sequence features among these HA antibodies were identified. Using this dataset, we further developed a memory B cell language model (mBLM) for antibody specificity prediction based on seven specificity categories, including HA head and stem domains. Saliency map explanation of mBLM revealed that key binding motifs were learned during specificity prediction. Moreover, we successfully applied mBLM to discover HA stem antibodies with subsequent experimental validation.

RESULTS

Examination of a large-scale collection of influenza antibodies reveals distinct features of HA head and stem antibodies

We compiled a list of 5,561 human monoclonal antibodies to influenza HA from 60 research publications and three patents (Table S1). Information on germline gene usage, sequence, binding specificity (e.g., group 1, group 2, type A or B, etc.), epitope (head or stem), and donor status (e.g., infected patient, vaccinee, etc.), if available, was collected for individual antibodies. Among these antibodies, which were isolated from 132 different donors, 565 (10.2%) bind to the globular head domain and 527 (9.5%) bind to the stem domain. Epitope information was not available for the remaining 4,469 HA antibodies.

We first aimed to analyze this large dataset to examine the recurring sequence features of human antibody responses to influenza HA. Our analysis captured previously known germline gene preference for HA stem antibodies, such as IGHV1-69[8,29] and IGHD3-9,[7] as well as for HA head antibodies, such as IGHV2-70 and IGHD4-17 (Figures 1A–1C and S1).[30] Other recurring sequence features were also observed in our analysis, such as the enrichment of IGKV3-11, IGKV3-15, and IGKV3-20 among HA stem antibodies, as well as IGKV1-33 and IGLV3-9 among HA head antibodies (Figure 1B). In addition, our analysis discovered five public clonotypes that target influenza type B HA (clonotypes 13, 16, 56, 89, and 117) that have not been described previously, to the best of our knowledge (Figures S2A and S2B; Table S1).

The high prevalence of IGHD4-17 among HA head antibodies stood out to us. It is known that the second reading frame of IGHD4-17 encodes a YGD motif (Figure S3A) and can pair with IGHV2-70 to form a multidonor antibody class targeting the receptor-binding site in the HA head domain.[30] However, our analysis here demonstrated that IGHD4-17 could pair with other IGHV genes to target diverse epitopes in the HA head domain

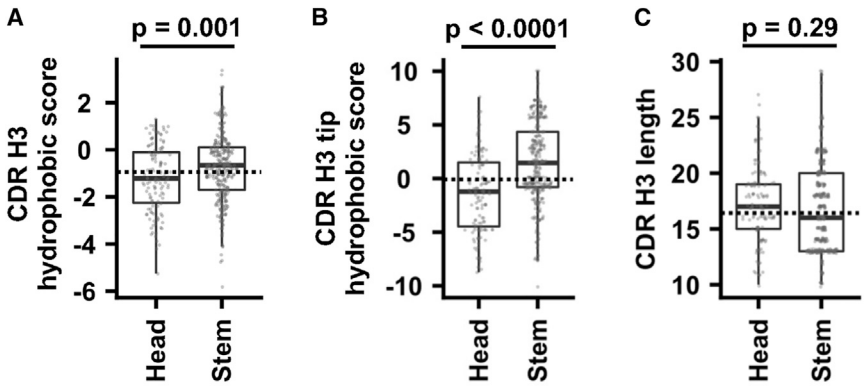


Figure 2. Hydrophobicity and length of CDR H3 sequences

(A and B) The hydrophobicity scores of (A) CDR H3 and (B) CDR H3 tip, as well as (C) the CDR H3 length, are compared between antibodies to HA head and HA stem domains. The *p* values were computed by two-tailed Student's t tests. For the boxplot, the middle horizontal line represents the median. The lower and upper hinges represent the first and third quartiles, respectively. The upper whisker extends to the highest data point within 1.5× inter-quartile range (IQR) of the third quartile, whereas the lower whisker extends to the lowest data point within 1.5× IQR of the first quartile. Each data point represents one antibody. The horizontal dotted line indicates the mean among antibodies from GenBank.

(Figures S3B and S3C; Table S2). Most of these antibodies contain an IGHD4-17-encoded YGD motif in the complementarity-determining region (CDR) H3 (Table S2). In fact, CDR H3 with a YGD motif was observed in 12.8% of the HA head antibodies but only in 0.8% and 2.0% of the HA stem antibodies and all antibodies from GenBank (Figure S3D; Table S3), respectively. These observations suggest the versatility of the IGHD4-17-encoded YGD motif in targeting multiple epitopes in the HA head domain, similar to the ability of IGHV3-53 to engage different epitopes in the SARS-CoV-2 spike (S) receptor-binding domain (RBD).[31,32]

Although the major antigenic sites in the HA head domain largely consist of hydrophilic and charged amino acids,[33–36] HA stem antibodies commonly target a hydrophobic groove.[37] Consistently, the CDR H3 sequences of HA stem antibodies had significantly higher hydrophobicity than those of HA head antibodies ($p = 0.001$) (Figure 2A). Such a difference was more pronounced when we only considered the tip of the CDR H3, which locates in the center of the CDR H3 sequence and is typically important for binding ($p < 0.0001$) (Figure 2B). By contrast, the CDR H3 lengths of antibodies to HA head and stem domains did not differ significantly ($p = 0.29$) (Figure 2C). Overall, these analyses reveal a prevalence for the YGD motif, specifically in HA head antibodies, and greater hydrophobicity in HA stem antibodies.

Antibody specificity prediction using mBLM

Due to the success of applying language models to predict protein structures and functions,[17–23] we postulated that antibodies with different specificities can be distinguished using a language model. Specifically, we aimed to pre-train a mBLM to learn the intrinsic "grammar" of functional antibodies and to subsequently distinguish between HA head and stem antibodies as well as antibodies to other antigens. Briefly, mBLM was pre-trained to predict masked amino acid residues in the context of paired heavy- and light-chain antibody sequences, using a total of 253,808 unique paired antibody sequences from GenBank[38] and Observed Antibody Space[39] (see STAR Methods). For antibody specificity prediction, mBLM was fine-tuned by using the final-layer embeddings of the pre-trained mBLM, followed by a multi-head self-attention block and a multi-layer perceptron (MLP) block (Figure 3A). Our prediction was based on seven specificity categories, namely influenza HA head, influenza HA stem, HIV, SARS-CoV-2 S NTD, SARS-CoV-2 S RBD, SARS-CoV-2 S S2, and others (none of the above). Because many antibodies in these specificity categories did not have light-chain sequences available, only heavy-chain sequences were used for specificity prediction (see STAR Methods). Training and test sets had a minimum pairwise Levenshtein distance of 10 and an average of 68 (Figure S4A). In other words, the pairwise sequence divergence between individual antibody sequences in the test set and the training set was at least 10 amino acid mutations, insertions, or deletions.

As indicated by the F1 score and confusion matrix analysis, mBLM had a decent performance on the test set (see STAR Methods; Figures 3B and 3C). In comparison, we also tested the performance of a k-nearest-neighbors (kNN) classifier, which was considered a baseline model, with varying values of k (1, 3, 5, 10, 20, 30, 50, 100, and 500) using the same training and test sets. Although the F1 score on the test set for the kNN classifier decreased from 0.68 to 0.53 as k increased from 1 to 500 (Figure S4B), the confusion matrix for the kNN classifier when k = 1 was quite poor (Figure S4C). Although the confusion matrix for the kNN classifier improved when k = 50 (Figure 3C), its F1 score on the test set was only 0.57, which was much lower than that of mBLM (F1 score on the test set = 0.79) (Figure 3B). We also fine-tuned a general protein language model ESM2[18] for antibody specificity prediction using the same training set. This fine-tuned ESM2 model (ESM-Ab) had an F1 score of 0.78 on the test set, which is comparable with mBLM (Figure 3B). Nevertheless, as compared with mBLM, ESM2-Ab had limited efficacy in distinguishing between SARS-CoV-2 NTD and RBD antibodies (Figure 3C). Furthermore, mBLM had only 41 million parameters, whereas ESM2 had 650 million,[18] demonstrating that mBLM is a more efficient and accurate model for antibody specificity prediction.

mBLM learned the sequence features of HA stem antibodies

Next, we aimed to understand what mBLM had learned for antibody specificity prediction. Advancements in the field of computer vision have employed gradient-weighted class activation maps (Grad-CAMs) on convolutional neural network (CNN)-based architectures to identify the determinants for classification decisions.[40,41] Here, Grad-CAM was adopted to analyze the fine-tuned mBLM by quantifying the importance of individual amino

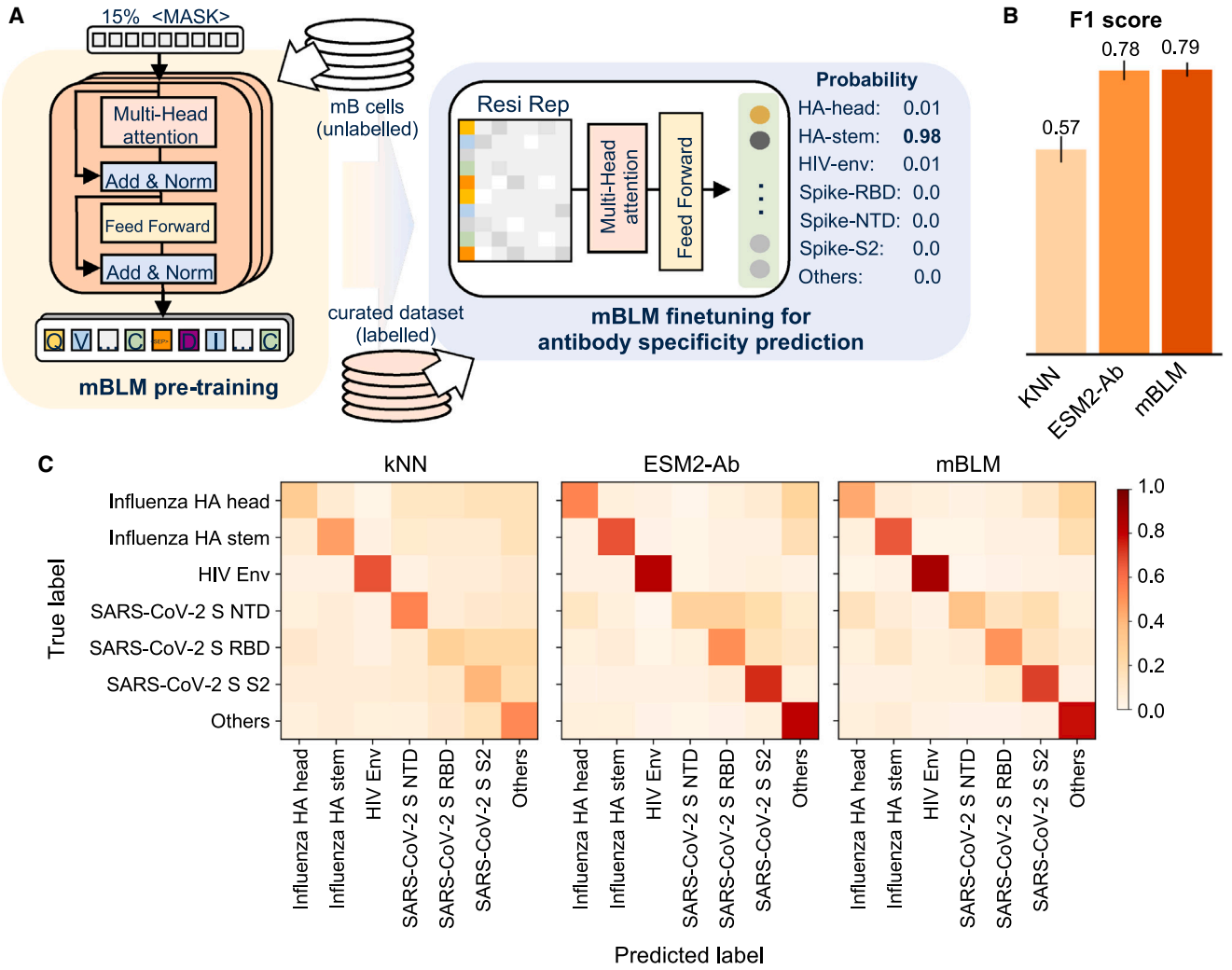


Figure 3. Antibody specificity prediction by mBLM

(A) Model architecture of memory B cell language model (mBLM) is shown. Arrows indicate the information flow in the network from the language model to antibody specificity prediction, with a final output of specificity class probability. Resi Rep, residual level representation (i.e., the final-layer embeddings from pre-trained mBLM).

(B) The performance of different antibody specificity prediction models was evaluated by F1 score, which represents the globally arithmetic mean of the harmonic means of precision and recall. Error bar represents standard deviation of 15-fold cross-validation. KNN: a baseline model using k-nearest neighbors algorithm. ESM2-Ab: pre-trained protein language model ESM2 was fine-tuned (same as mBLM) for antibody specificity prediction.[18]

(C) Model performance of mBLM on the test set was evaluated by a normalized confusion matrix.

acid residues for antibody specificity prediction (Figure 4A), where a higher saliency score represents higher importance.

Based on the saliency score pattern, we further identified six clusters of HA stem antibodies. These clusters captured several known sequence features of HA stem antibodies. For example, most antibodies in cluster 3 were encoded by IGHD3-9 (Figure 4B), which is known to be enriched among HA stem antibodies (Figure 1C).[7] Among IGHD3-9 antibodies in cluster 3, we observed an FxWL motif in the CDR H3 with high saliency score (Figure 4C). As described previously, many IGHD3-9 antibodies are featured by a LxYFxWL motif in the CDR H3.[7] Therefore, our result indicates that the fine-tuned mBLM partially learned a known CDR H3 motif for predicting HA stem antibodies. Other known sequence features of HA stem antibodies were also learned by mBLM, including IGHV1-18 with a QxxV motif in the CDR H3 (Figures S5A and S5B),[43] IGHV1-69 with Y98 (Figures S5A–S5D),[8] and IGHV6-1 with an FGV motif in the CDR H3 (Figures S5E and S5F).[54] For antibody residues, the Kabat numbering scheme is used unless otherwise stated.

When we projected the saliency score of individual residues on the structures, residues closer to the epitope appeared to have a higher saliency score (Figures 4D and S5G–S5I). Consistently, through systematically analyzing 18 structures of HA stem antibodies,[7,29,42–53] we found that the saliency score of individual residues in HA stem antibodies and their distance to HA exhibited a moderate negative correlation (Spearman's rank correlation = −0.38, Figure 4E). This result can be at least partly explained by the enrichment of residues with high saliency score

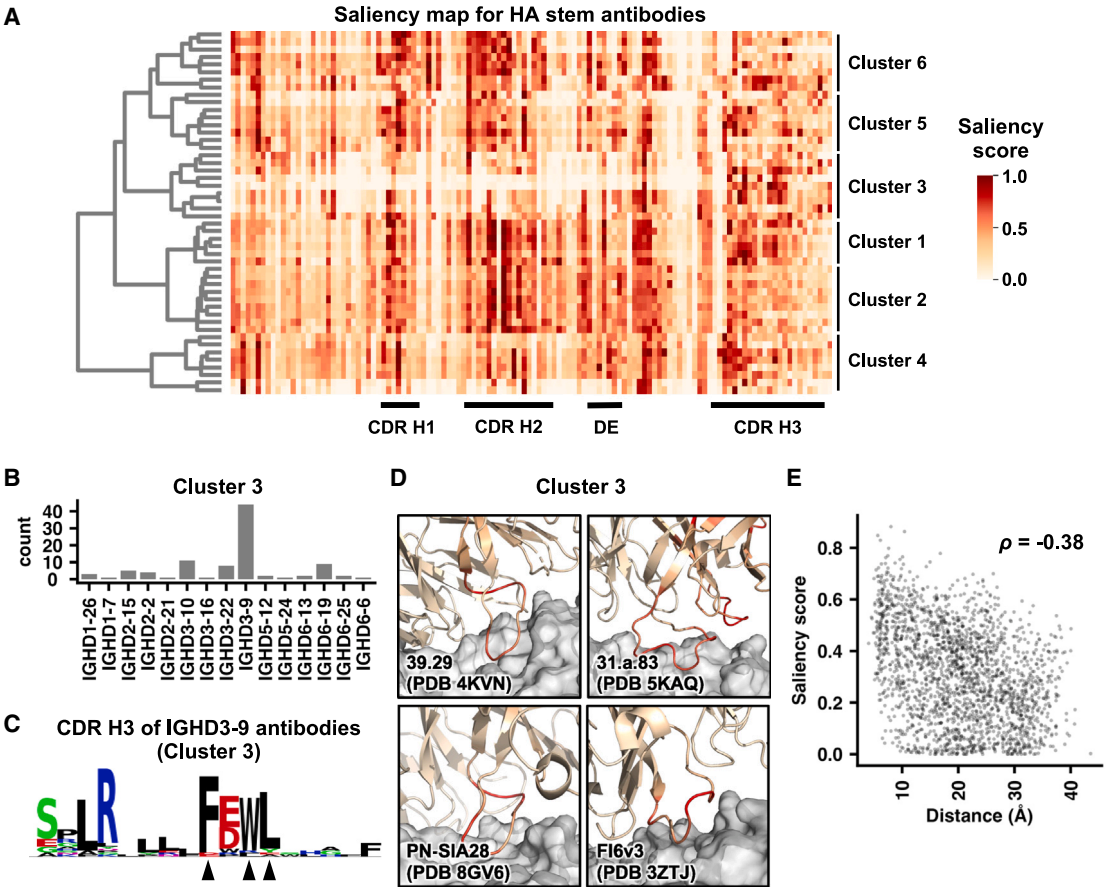


Figure 4. Explanation of mBLM using saliency score

(A) Saliency score for each residue in individual HA stem antibodies was shown as a heatmap. Each row represents a single HA stem antibody. X axis represents the amino acid residue of the heavy chain. Regions corresponding to CDR H1, H2, H3, and DE loop are indicated. For visualization purpose, only 50 HA stem antibodies are shown. Six clusters of HA stem antibodies were identified using hierarchical clustering with Ward's method.

(B) IGHD gene usage among antibodies in cluster 3 is shown.

(C) The saliency score of each CDR H3 residue in IGHD3-9 antibodies within cluster 3 was analyzed. The frequency of each amino acid for residues with a saliency score >0.5 is shown as a sequence logo. Arrows at the bottom indicate the residues of interest.

(D) Saliency scores are projected onto the structures of four antibodies in cluster 3, namely 39.29 (PDB: 4KVN),[42] 31.a.83 (PDB: 5KAQ),[43] PN-SIA28 (PDB: 8GV6),[44] and FI6v3 (PDB: 3ZTJ).[45] The color scheme is the same as that in (A).

(E) The relationship between saliency score and distance to the antigen (i.e., HA stem) is shown as a scatterplot. Spearman's rank correlation coefficient (ρ) is indicated. A total of 18 structures of HA stem antibodies in complex with HA were analyzed (PDB: 3FKU, 3GBN, 3SDY, 3ZTJ, 4FQI, 4KVN, 4NM8, 4R8W, 5JW3, 5KAN, 5KAQ, 5K9K, 5K9O, 5K9Q, 5WKO, 6E3H, 6NZ7, and 8GV6).[7,29,42–53]

in the CDRs (Figure 4A), which are closer to the binding interface and information-rich due to their high sequence diversity.

To gain additional understanding of mBLM, we analyzed the final-layer embeddings of the pre-trained mBLM using t-distributed stochastic neighbor embedding (t-SNE). Specifically, heavy-chain sequences in the training set for fine-tuning were projected into a two-dimensional space according to the embeddings. The result showed clustering of antibodies that belonged to the same V gene family (Figure S4D). Moreover, distinct clusters can be observed for antibodies from the same specificity category (Figure S4E) as well as different HA stem antibody clusters (Figures 4A and S4F). These observations demonstrated that antibodies with shared sequence features were proximal in the embedding space, although it may be expected because a similar observation was made when antibody sequences were directly embedded.[55]

Sequence determinants of an IGHV1-46 HA stem antibody

Given that the saliency score analysis was able to capture known sequence features of HA stem antibodies, we postulated that it could also be applied to identify previously unknown sequence features. Here, we applied the saliency score analysis to the HA stem antibody C1-3.7F02, which was isolated from the plasmablasts of a healthy donor after receiving a trivalent seasonal vaccine and shown to bind to H3 HA.[56] Because there were only 12 known IGHV1-46 HA stem antibodies in the dataset that we assembled (Table S1), we did not expect mBLM to be well trained for identifying HA stem antibodies that were encoded by IGHV1-46. Yet, C1-3.7F02 had a relatively high confidence score of 0.64 in the test set (range from 0 to 1, 1 being the highest confidence). To understand the sequence determinants of C1-3.7F02, we analyzed the saliency score of each residue of C1-3.7F02 heavy

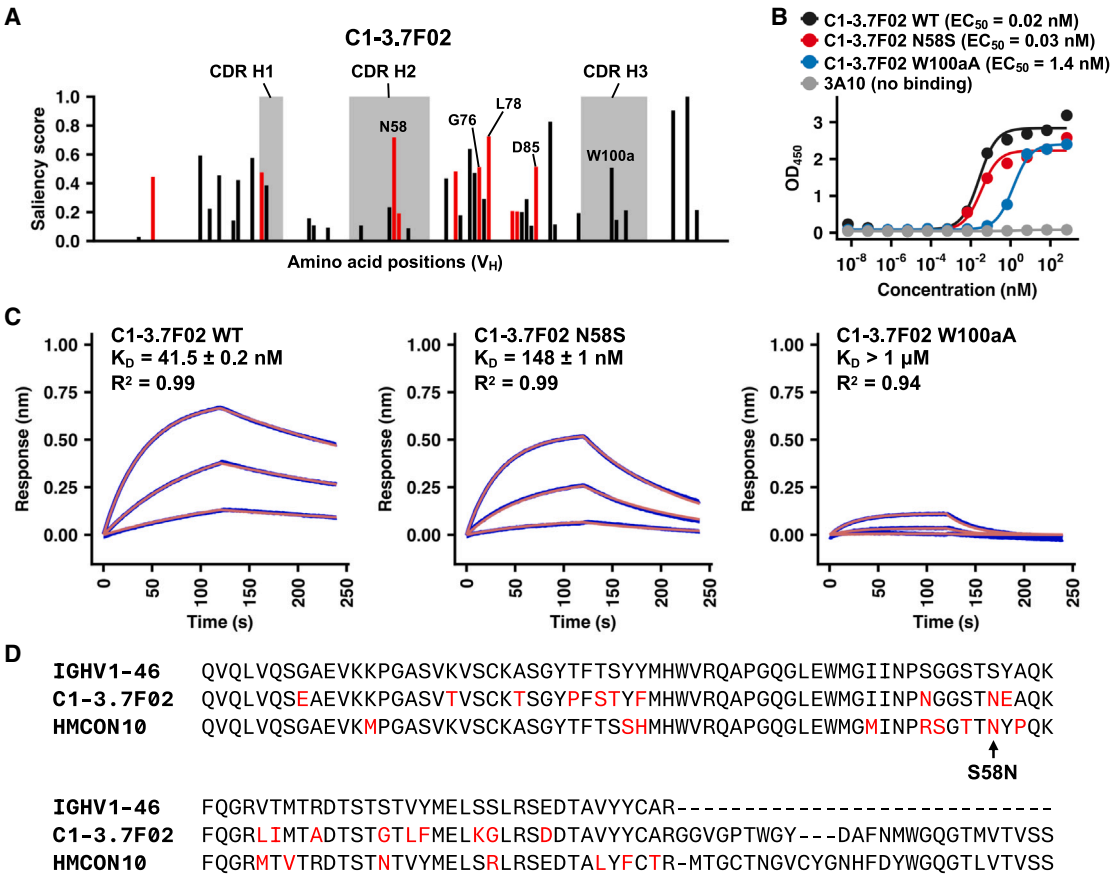


Figure 5. Sequence determinants for the HA-stem-binding activity of C1-3.7F02

(A) Saliency score of each residue of C1-3.7F02[56] is shown as a bar chart. Residues that represent somatic hypermutations are colored in red. Residues of interest are labeled.

(B) The binding affinity of C1-3.7F02 wild type (WT) (black), N58S mutant (red), and W100aA mutant (blue) IgGs against H3 mini-HA was measured by ELISA. Their EC_{50} values are indicated. 3A10 is an influenza neuraminidase antibody and serves as a negative control here.[58]

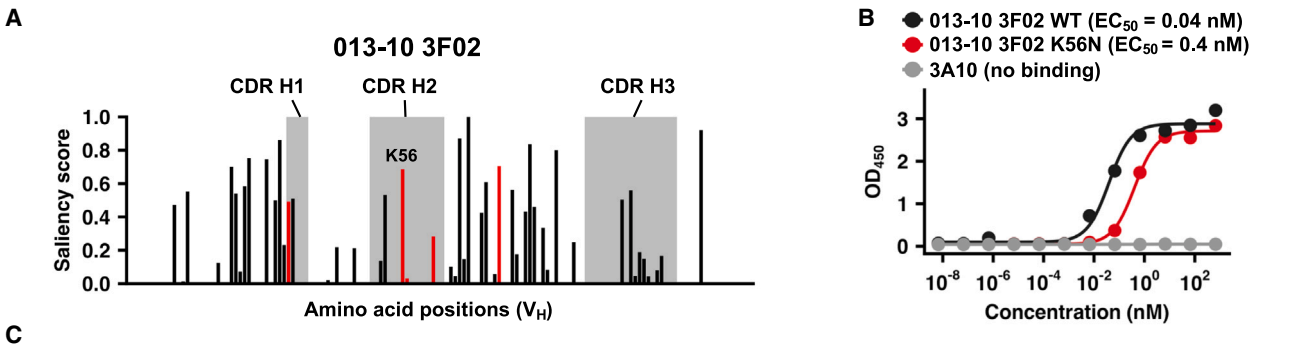
(C) Binding kinetics of different Fabs against recombinant H3 mini-HA[57] were measured by biolayer interferometry (BLI). The y axis represents the response. Blue lines represent the response curves and red lines represent a 1:1 binding model. Binding kinetics were measured for four concentrations of Fab at 3-fold dilution, ranging from 300 to 33 nM. Dissociation constant (K_D) and the goodness of model fitting (R^2) are indicated.

(D) Sequence alignment of IGHV1-46 germline sequence with the heavy-chain sequences of C1-3.7F02[56] and HMCON10[59] was performed using MAFFT.[60] Residues that represent somatic hypermutations in the V gene are colored in red.

chain (Figure 5A). Several residues in the CDRs of C1-3.7F02 had high saliency scores, including N58 in CDR H2 and W100a in CDR H3. N58 represented a somatic hypermutation because the IGHV1-46 germline-encoded amino acid at residue 58 is Ser (Figure 5D). Germline reversion mutation N58S weakened the EC_{50} of C1-3.7F02 immunoglobulin G (IgG) to H3 mini-HA by ~50% (Figure 5B) and the K_D of C1-3.7F02 Fab to H3 mini-HA, which was an HA-stem-only construct designed based on human H3N2 A/Finland/486/2004 HA,[57] by >3-fold (Figure 5C). Our result indicates that somatic hypermutation S58N is important for the affinity maturation of C1-3.7F02 against HA stem. To probe for the importance of W100a in HA-stem binding, mutation W100aA was introduced. W100aA reduced the binding of both C1-3.7F02 IgG and Fab by at least one order of magnitude in terms of EC_{50} and K_D, respectively. This result indicates that W100a is also a key residue for HA-stem binding.

Besides C1-3.7F02, there was only one other IGHV1-46 antibody targeting group 2 HA stem in our dataset, namely HMCON10[59] (Table S1). C1-3.7F02 and HMCON10 were isolated from different donors.[56,59] Although both C1-3.7F02 and HMCON10 had a relatively large number of somatic hypermutations in the heavy-chain V gene (19 and 16 amino acids, respectively), S58N was the only common somatic hypermutation between them (Figure 5D). This observation further implies that S58N plays a role in the affinity maturation of other IGHV1-46 antibodies against group 2 HA stem.

Several somatic hypermutations in the framework region 3 of C1-3.7F02 also had high saliency scores, with the top three being G76, L78, and D85 (Figure 5A). To test their importance for HA-stem binding, germline reversion mutations were introduced at these three residues, namely G76S, L78V, and D85E. Both G76S and L78V weakened the K_D by 2-fold, whereas D85E had a milder impact on K_D (Figure S6). In fact, G76 and L78 locate within the heavy-chain DE loop (residues 71–78), which is often referred to as CDR H4 and sometimes involves in antigen binding.[61] Together, our findings indicate that somatic hypermutations in framework



C

```
IGHV3-30       QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYGMHWVRQAPGKGLEWVAVISYDGSNKYYADSVKGRFT
013-10 3F02    QVQLVESGGGVAQPGRSLRLSCAASGFTFSTYGMHWVRQAPGKGLEWVAVISYDGSKRYYAEILKDRFT
3I14           QVQLLESGGGVVQPGRSLRLSCAASGFTFSNYGMHWVRQAPGKGLEWVAIISFDGSKKYYANSVKGRST
FI3082V        QVRLVESGGGVVQPGRSLRLSCAASGFTFSNFGMNWVRQAPGKGLEWVAVISYDGSKKFYADSVNSRFT
310-18C10      QVHLVESGGGVVQPGRSQRLSCAASGVTFSMYTMHWVRQAPGKGLEWVAVISYDGSKKDYADSVKGRFT
81.39          EVQLVESGGGVVQPGRSLRLSCAASGFAFHNRAMHWVRQAPGKGLEWVALIYFDGSKQYYADSVKGRFT
                                                                  N56K

IGHV3-30       ISRDNSKNTLYLQMNSLRAEDTAVYYCAK----------------------------------------
013-10 3F02    ISRDNSKNTVYLQMNSLRTEDTAVYFCAK-------EGRPLIVSGTHFFQFYYGMDVWGQGTTVTVSS
3I14           ISRDNSKNTLSLQMNSLGPEDTALYYCAKLPSPYYFDSRFVWVAASAF-------HFWGQGILVTVSS
FI3082V        ISRDNSKNTLFLQMNTLRPEDTATYYCAK-------DQTVVSVAAALF-------DYCGQGTLVTVSS
310-18C10      ISRDNSKTTMYLQMNSLRAEDTAVYFCAR-------D------VGGYFQRSYF--DSWGQGTLVTVSS
81.39          ISRDNSKNTVFLQMNSLRPEDTAVYYCAV-PGPIF----GIFPPWSYF-------DHWGQGILVTVSS
```

Figure 6. Sequence determinants for the HA-stem-binding activity of 013-10 3F02

(A) Saliency score of each residue of 013-10 3F02[62] is shown as a bar chart. Residues that represent somatic hypermutations are colored in red. Residue 56 is labeled.

(B) The binding affinity of 013-10 3F02 WT (black) and K56N mutant (red) IgGs against H1 mini-HA was measured by ELISA. Their EC_{50} values are indicated. 3A10 is an influenza neuraminidase antibody and serves as a negative control here.[58]

(C) Sequence alignment of IGHV3-30 germline sequence with the heavy-chain sequences of 013-10 3F02,[62] 3I14,[64] FI3082,[65] 310-18C10,[42] and 81.39[42] was performed using MAFFT.[60] Residues that represent somatic hypermutations in the V gene are colored in red.

region 3, particularly the DE loop, contribute to the affinity maturation of C1-3.7F02.

Sequence determinants of an IGHV3-30 HA stem antibody

We further performed a saliency score analysis on another HA stem antibody 013-10 3F02, which was encoded by IGHV3-30 and shown to bind to group 1 HA.[62] Similar to C1-3.7F02, 013-10 3F02 was also isolated from the plasmablasts of a healthy donor after receiving a trivalent seasonal vaccine[62] and had a decent confidence score of 0.54 in the test set. Although IGHV3-30 was the third most commonly used IGHV gene among HA stem antibodies (7.6%, Figure 1A), the sequence determinants of IGHV3-30 HA stem antibodies are not as well characterized as those encoded by IGHV1-69 (34% among HA stem antibodies) and IGHV1-18 (6.0% among HA stem antibodies).[8,43] K56 of 013-10 3F02, which represented a somatic hypermutation in the CDR H2, had a high saliency score (Figure 6A). Reverting K56 to the IGHV3-30 germline-encoded N56 weakened the binding affinity of 013-10 3F02 to H1 mini-HA, which was an HA-stem-only construct designed based on human H1N1 A/Brisbane/59/2007 HA,[63] by 10-fold (Figure 6B). Our dataset contained 64 IGHV3-30 HA stem antibodies from 17 different donors (Table S1). Most of these IGHV3-30 HA stem antibodies (39/64) were encoded by IGHD3-9, which can result in a CDR H3-dominant binding mode that is largely independent of IGHV gene usage.[7] Although none of the 39 IGHV3-30/IGHD3-9 HA stem antibodies contained N56K, it was observed in 5 out of 25 (20%) IGHV3-30/non-IGHD3-9 HA stem antibodies (Figure 6C). These five antibodies, including 013-10 3F02, were isolated from five out of seven donors who had IGHV3-30/non-IGHD3-9 HA stem antibodies. The enrichment of N56K among IGHV3-30/non-IGHD3-9 HA stem antibodies suggests that N56K facilitates their affinity maturation. Together with the analysis above on C1-3.7F02, our results suggest that mBLM can help identify somatic hypermutations that are critical for affinity maturation.

Discovering HA stem antibodies using mBLM

There are two non-overlapping epitopes in the HA stem, namely central stem epitope[46,47] and anchor stem epitope.[66,67] After we had assembled our HA antibody dataset (Table S1), a study reported 60 HA antibodies to the central stem epitope and 38 to the anchor stem epitope.[68] Although these antibodies were not in the HA antibody dataset that we assembled (Table S1), they provided an additional opportunity to test the fine-tuned mBLM. Among the 60 antibodies to the central stem epitope,

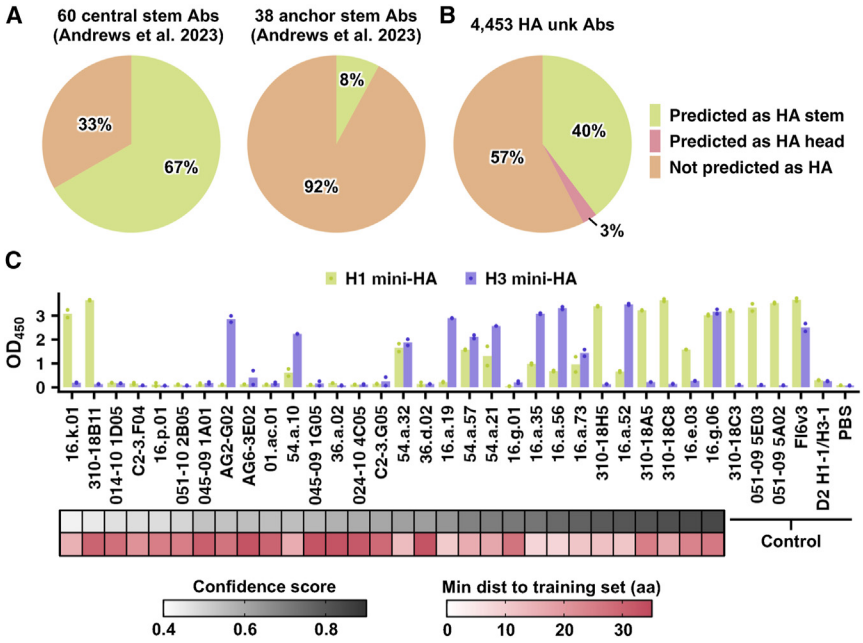


Figure 7. Discovery of HA stem antibodies by mBLM

(A and B) mBLM was applied to predict the specificity of (A) 60 antibodies to central stem epitope (left) and 38 to anchor stem epitope (right) that have been reported,[68] as well as (B) 4,453 HA antibodies with unknown epitopes (HA unk) in the dataset that we assembled. The fraction of antibodies that were predicted to bind to HA stem domain (predicted as HA stem), HA head domain (predicted as HA head), or to other antigens (not predicted as HA) is shown. (C) Using ELISA, the binding of 30 HA unk antibodies that were predicted as HA stem antibodies was tested against H1 mini-HA[63] and H3 mini-HA,[57] both of which were HA-stem-only constructs. The confidence score of each of these antibodies as HA stem antibodies as well as their sequence divergence to the most similar antibodies in the training set (min dist to training set) are shown as heatmaps. Four known HA stem antibodies (051-09 5A02, 051-09 5E03, 310-18C3, and FI6v3)[45,62,69] were included as positive control. D2 H1-1/H3-1, which is a known HA head antibody,[70] was included as a negative control. In this binding experiment, antibodies were not purified from the supernatant and thus their concentrations were unknown.

the fine-tuned mBLM correctly predicted 67% (40/60) as HA stem antibodies (Figure 7A). By contrast, among the 38 antibodies to the anchor stem epitope, only 8% (3/38) were predicted as HA stem antibodies (Figure 7A). The poor performance of the fine-tuned mBLM on antibodies to anchor stem epitope was likely due to lack of antibodies to anchor stem epitope in the dataset that we assembled (Table S1). In fact, antibodies to anchor stem epitope have only been extensively characterized in the last two years.[67] These results suggest that HA stem antibodies correctly predicted by mBLM would mostly target the central stem epitope.

Among the 5,561 HA antibodies in the dataset that we assembled (Table S1), 80% (4,469/5,561) have unknown epitopes, of which 4,453 have heavy-chain sequence information available. Subsequently, we applied the fine-tuned mBLM to predict the specificities of these 4,453 antibodies. Although 40% (1,767/4,453) were predicted as HA stem antibodies, only 3% (119/4,453) were predicted as HA head antibodies (Figure 7B). Consistently, many antibodies in our dataset came from studies that would, by design, result in the enrichment of HA stem antibodies during antibody isolation.[43,45,46,59,65,69,71,72] However, this bias could not fully explain the poor performance of the fine-tuned mBLM on HA head antibodies. Unlike the highly conserved HA stem domain,[28] HA head domain has a huge sequence diversity across influenza strains and subtypes. Additionally, HA head domain has more antibody-binding sites than the HA stem domain.[37] Consequently, HA head antibodies were expected to have a much higher sequence diversity than HA stem antibodies. The poor performance of the fine-tuned mBLM on HA head antibodies was likely due to insufficient sequences for each specific binding site for HA head antibodies in our training set (see discussion).

To experimentally validate our prediction result, 30 antibodies that were predicted to target HA stem, spanning a range of confidence scores (0.43 to 0.84), were individually expressed and tested for binding to stem-only constructs, namely H1 mini-HA and H3 mini-HA.[57,63] H1 mini-HA represented group 1 HA stem, whereas H3 mini-HA represented group 2 HA stem. We only include antibodies that are known to bind influenza A HA in this validation experiment. Antibodies that are known to target influenza B HA were excluded. Our enzyme-linked immunosorbent assay (ELISA) result showed that 57% (17/30) could bind to either H1 mini-HA or H3 mini-HA or both (Figure 7C). This validation rate appeared to correlate with the confidence score of the model. Among the 16 antibodies with confidence scores > 0.6, 13 (81%) were validated as HA stem antibodies. By contrast, among the remaining 15 antibodies with confidence scores < 0.6, only four (29%) were validated as HA stem antibodies. For the 17 validated HA stem antibodies, 11 were encoded by IGHV1-18 with QxxV motif in the CDR H3 (confidence scores range from 0.55 to 0.84), which is a known sequence feature of HA stem antibodies.[43] Five were encoded by IGHV1-69 with Y98 (confidence scores range from 0.43 to 0.83), which is another known sequence feature of HA stem antibodies.[8] The remaining antibody, AG2-G02 (confidence score = 0.54), was encoded by IGHV1-2.

AG2-G02 was originally isolated from the plasmablasts of a healthy donor after receiving a trivalent seasonal vaccine.[56] In other words, AG2-G02 did not come from a study that aimed to isolate HA stem antibodies. Unlike the other 16 validated HA stem antibodies, AG2-G02 did not contain any well-characterized sequence feature of HA stem antibodies. AG2-G02 has been shown to target human H3N2 HA, with a preference toward older strains.[56] Consistently, our data showed that AG2-G02 bound strongly to H3 mini-HA but not H1 mini-HA (Figures 7C, S7A, and S7B). We also showed that AG2-G02 cross-reacted with an H3N8 HA but not a more recent human H3N2 HA and other group 2 HAs tested (Figure S7A). Overall, these results

demonstrate that the fine-tuned mBLM enables discovery of antibodies to known epitopes.

Characterization of an IGHV1-69 HA stem antibody

IGHV1-69 HA stem antibodies have diverse CDR H3 sequences.[8,73] Although many IGHV1-69 HA stem antibodies are featured by a Tyr at residue 98, it is neither a necessary nor a sufficient determinant of HA-stem-binding activity.[8,73] For example, some IGHV1-69 HA stem antibodies do not even have a Tyr in the CDR H3.[8] Despite the high sequence diversity of IGHV1-69 HA stem antibodies, mBLM was able to predict certain IGHV1-69 HA stem antibodies with high confidence, as exemplified by 310-18A5[69] (confidence score = 0.81), which has a Levenshtein distance of 25 (i.e., ~21% sequence divergence) to the most similar antibody in the training set. Biolayer interferometry indicated that 310-18A5 had a strong binding affinity against the HA from H1N1 A/Solomon Island/3/2006 (K_D = 0.2 nM, Figure S7C) as well as H1 mini-HA (K_D = 1.0 nM, Figure S7D). Besides, 310-18A5 had neutralization activity against two antigenically distinct H1N1 strains (Figure S7E). Consistently, cryo-EM analysis confirmed that 310-18A5 bound to the HA stem domain with a similar approaching angle as other IGHV1-69 HA stem antibodies that have structural information available (Figures S7F and S7G; Table S4).[29,46,47,49] This observation substantiates that IGHV1-69 HA stem antibodies, albeit with high sequence divergence, converge to similar binding modes.

DISCUSSION

Although influenza HA antibodies have been studied over decades, there has been a lack of effort to summarize information about these antibodies. In this study, we performed a large-scale analysis of more than 5,000 influenza HA antibodies by mining research publications and patents. Although many recurring sequence features of influenza HA antibodies are reported in individual studies,[7,8,29,30,43,54,67] our results revealed additional ones that have not been described, to the best of our knowledge. For example, our study discovered the enrichment of YGD motif in the CDR H3 of HA head antibodies as well as multiple public clonotypes to influenza type B HA. We further developed a language model for antibody specificity prediction, which was subsequently applied to reveal sequence determinants of HA stem antibodies and discover HA stem antibodies. Overall, this work not only advances the molecular understanding of influenza HA antibodies but also provides an important resource for the antibody research community (Table S1).

Discovering antibodies to a specific antigen of interest typically requires less effort than epitope mapping. Consistently, epitope information (head or stem) is available for only ~20% of HA antibodies in our dataset. Nevertheless, we were able to utilize these ~20% of HA antibodies to train mBLM to identify HA stem antibodies among the remaining ~80% with no epitope information. This result demonstrates that mBLM can accelerate epitope mapping. A potential application of mBLM is to map the epitopes of antibodies that are discovered from single-cell B cell receptor sequencing of plasmablasts or antigen-specific memory B cells. We believe that mBLM synergizes with existing high-throughput antibody discovery approaches to streamline the analysis of antibody responses. Although our work here applied mBLM to predict antibody specificity based on seven specificity categories, it can be fine-tuned to extend to any specificities, as long as sufficient and diverse antibody sequences with such specificities are available. In this respect, the continuous improvement of the speed of antibody discovery and characterization will also be beneficial, if not essential.[3,4] Given that many antibodies with different specificities are characterized in the literature, future generalization of mBLM to additional antibody specificities will likely be achievable by extensive data mining.

Similar to any machine learning algorithm, a key requirement for mBLM to make accurate predictions is to have a large dataset for training. Therefore, an important question is: how many antibodies are needed for model training? The answer may vary among specificity categories. For example, mBLM prediction of antibodies to HA stem (central stem in particular) worked quite well with around 500 HA stem antibodies in the training set. Conversely, the performance of mBLM on HA head antibodies was less successful, although the number of HA head antibodies in our training set was only ~25% less than that of HA stem antibodies. This difference indicates that accurate prediction of HA head antibodies by mBLM will require a lot more HA head antibodies in the training set. One possible explanation is that the HA head domain has more antibody-binding sites than the HA stem domain. Although most antibodies against the HA stem domain target the central stem epitope and the more recently discovered anchor stem epitope,[66,67] antibodies against the HA head domain target not only the five major antigenic sites but also lateral patch, receptor-binding site, vestigial esterase subdomain, and trimer interface.[37] The larger number of antibody-binding sites in the HA head domain would lead to a higher sequence diversity of HA head antibodies. Based on our prediction results of HA central stem antibodies, we estimate that at least 500 antibodies to a specific antibody-binding site are needed for training mBLM to achieve accurate prediction. Accordingly, accurate prediction of HA head antibodies will require at least a few thousands of HA head antibodies in the training set. If we further account for the hypervariability of the HA head domain across HA subtypes, the number of HA head antibodies required would most likely be a magnitude higher.

As more antibody sequences become available for different HA antibody-binding sites, future versions of mBLM may provide a finer epitope classification of the HA. For example, instead of treating HA head domain as a single specificity category, different antibody-binding sites within the HA head domain could each be classified as their own specificity categories. However, fine epitope mapping of a large number of antibodies is unlikely to be realized in the short term due to the tedious experimental efforts required. In comparison, mapping antibody specificity to a protein domain, such as the HA head domain, is more practical experimentally. As demonstrated by our saliency score analysis of HA stem antibodies, mBLM is able to identify subcategories of antibodies within a given specificity category. Thus, binning different antibody-binding sites within a protein domain as a single specificity category provides a short-term solution until sufficient antibody sequences for a given antibody-binding site are available.

The success of applying a deep learning model to protein research can largely be attributed to the presence of databases

such as Protein Data Bank (PDB),[74] UniProt,[75] and UniRef,[76] which describe the sequence-structure-function relationships. Similarly, most, if not all, existing models for antibody specificity prediction were trained using structural information of antibody-antigen interactions in PDB.[16] Nevertheless, the epitopes of most antibodies in the literature are mapped by non-structural approaches, such as competition or mutagenesis experiments.[77] These epitope mapping data, despite being obtained by non-structural approaches, are tremendously useful for training a model for antibody specificity prediction, as shown by our study here. Consequently, future efforts should focus on establishing a centralized database that describes the sequence-specificity relationship for antibodies, even for those without structural information available. Such database will allow the power of deep learning models to be fully harnessed in antibody research.

Limitations of the study

We acknowledge that some antibodies are polyreactive or polyspecific,[78] as exemplified by the antibody 2G12, which cross-reacts with HIV envelope,[79] influenza HA,[80] and SARS-CoV-2 S.[81] A limitation of the current model architecture of mBLM is that it did not account for polyreactivity or polyspecificity. A potential solution is to modify the model architecture for multi-label classification[82] to predict polyreactivity or polyspecificity. Similarly, multi-label classification should also enable prediction of antibody breadth. However, most of the known antibodies have not been tested for polyreactivity or polyspecificity. Although antibody breadth is more commonly tested, the panels of viral strains used in different studies are almost always different. As a result, the major barrier for adopting multi-label classification for antibody specificity prediction is the limited data availability. Another limitation of mBLM is that it does not take biophysical knowledge into account. If the underlying physical laws are also encoded by the model, the amount of data required for accurate prediction could be substantially reduced.[82,83] Thus, the performance of mBLM should improve by incorporating antibody-antigen structural information into model training, especially for specificity categories with limited known antibody sequences. Consistently, a previous study shows that sequence features derived from the structural analysis of a single antibody-antigen complex facilitate discovery of HA head antibodies.[84] Because many antibody-antigen complex structures are publicly available, development of a biophysics-informed mBLM is warranted in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.immuni.2024.07.022.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) DP2 AT011966 (N.C.W.), R01 AI167910 (N.C.W.), the Michelson Prizes for Human Immunology and Vaccine Research (N.C.W.), the Searle Scholars Program (N.C.W.), and Carl R. Woese Institute for Genomic Biology Postdoctoral Fellowship (H.L.). We thank Kristen Flatt at the UIUC Materials Research Laboratory Central Research Facilities for assistance with cryo-EM experiments, as well as Meng Yuan and Zongjun Mou for helpful discussions.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study. Y.W., H.L., and N.C.W. assembled the dataset. Y.W., Y.-H.Y., and N.C.W. performed data analysis. H.L., Q.W.T., A.B.G., W.O.O., T.J.C.T., D.C., and I.R.S. performed the antibody-binding experiments. R.L., C.S.G., and X.C. purified the proteins and performed the cryo-EM analysis. Y.W., H.L., R.L., and N.C.W. wrote the paper and all authors reviewed and/or edited the paper.

DECLARATION OF INTERESTS

N.C.W. consults for HeliXon.

Received: October 18, 2023
Revised: April 24, 2024
Accepted: July 24, 2024
Published: August 19, 2024

REFERENCES

1. Graham, B.S., Gilman, M.S.A., and McLellan, J.S. (2019). Structure-based vaccine antigen design. Annu. Rev. Med. *70*, 91–104. https://doi.org/10.1146/annurev-med-121217-094234.
2. Lu, R.M., Hwang, Y.C., Liu, I.J., Lee, C.C., Tsai, H.Z., Li, H.J., and Wu, H.C. (2020). Development of therapeutic antibodies for the treatment of diseases. J. Biomed. Sci. *27*, 1. https://doi.org/10.1186/s12929-019-0592-z.
3. Winters, A., McFadden, K., Bergen, J., Landas, J., Berry, K.A., Gonzalez, A., Salimi-Moosavi, H., Murawsky, C.M., Tagari, P., and King, C.T. (2019). Rapid single B cell antibody discovery using nanopens and structured light. mAbs *11*, 1025–1035. https://doi.org/10.1080/19420862.2019.1624126.
4. Curtis, N.C., and Lee, J. (2020). Beyond bulk single-chain sequencing: getting at the whole receptor. Curr. Opin. Syst. Biol. *24*, 93–99. https://doi.org/10.1016/j.coisb.2020.10.008.
5. Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody

repertoire. Nature *566*, 393–397. https://doi.org/10.1038/s41586-019-0879-y.

6. Schroeder, H.W., Jr. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Dev. Comp. Immunol. *30*, 119–135. https://doi.org/10.1016/j.dci.2005.06.006.
7. Wu, N.C., Yamayoshi, S., Ito, M., Uraki, R., Kawaoka, Y., and Wilson, I.A. (2018). Recurring and adaptable binding motifs in broadly neutralizing antibodies to influenza virus are encoded on the D3-9 segment of the Ig gene. Cell Host Microbe *24*, 569–578.e4. https://doi.org/10.1016/j.chom.2018.09.010.
8. Avnir, Y., Tallarico, A.S., Zhu, Q., Bennett, A.S., Connelly, G., Sheehan, J., Sui, J., Fahmy, A., Huang, C.Y., Cadwell, G., et al. (2014). Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. PLoS Pathog. *10*, e1004103. https://doi.org/10.1371/journal.ppat.1004103.
9. Zhou, T., Lynch, R.M., Chen, L., Acharya, P., Wu, X., Doria-Rose, N.A., Joyce, M.G., Lingwood, D., Soto, C., Bailer, R.T., et al. (2015). Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. Cell *161*, 1280–1292. https://doi.org/10.1016/j.cell.2015.05.007.
10. Robbiani, D.F., Bozzacco, L., Keeffe, J.R., Khouri, R., Olsen, P.C., Gazumyan, A., Schaefer-Babajew, D., Avila-Rios, S., Nogueira, L., Patel, R., et al. (2017). Recurrent potent human neutralizing antibodies to Zika virus in Brazil and Mexico. Cell *169*, 597–609.e11. https://doi.org/10.1016/j.cell.2017.04.024.
11. Ehrhardt, S.A., Zehner, M., Krähling, V., Cohen-Dvashi, H., Kreer, C., Elad, N., Gruell, H., Ercanoglu, M.S., Schommers, P., Gieselmann, L., et al. (2019). Polyclonal and convergent antibody response to Ebola virus vaccine rVSV-ZEBOV. Nat. Med. *25*, 1589–1600. https://doi.org/10.1038/s41591-019-0602-4.
12. Cohen-Dvashi, H., Zehner, M., Ehrhardt, S., Katz, M., Elad, N., Klein, F., and Diskin, R. (2020). Structural basis for a convergent immune response against Ebola virus. Cell Host Microbe *27*, 418–427.e4. https://doi.org/10.1016/j.chom.2020.01.007.
13. Chen, E.C., Gilchuk, P., Zost, S.J., Suryadevara, N., Winkler, E.S., Cabel, C.R., Binshtein, E., Chen, R.E., Sutton, R.E., Rodriguez, J., et al. (2021). Convergent antibody responses to the SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Rep. *36*, 109604. https://doi.org/10.1016/j.celrep.2021.109604.
14. Claireaux, M., Caniels, T.G., de Gast, M., Han, J., Guerra, D., Kerster, G., van Schaik, B.D.C., Jongejan, A., Schriek, A.I., Grobben, M., et al. (2022). A public antibody class recognizes an S2 epitope exposed on open conformations of SARS-CoV-2 spike. Nat. Commun. *13*, 4539. https://doi.org/10.1038/s41467-022-32232-0.
15. Wang, Y., Yuan, M., Lv, H., Peng, J., Wilson, I.A., and Wu, N.C. (2022). A large-scale systematic survey reveals recurring molecular features of public antibody responses to SARS-CoV-2. Immunity *55*, 1105–1117.e4. https://doi.org/10.1016/j.immuni.2022.03.019.
16. Cia, G., Pucci, F., and Rooman, M. (2023). Critical review of conformational B-cell epitope prediction methods. Brief. Bioinform. *24*, bbac567. https://doi.org/10.1093/bib/bbac567.
17. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA *118*, e2016239118. https://doi.org/10.1073/pnas.2016239118.
18. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.
19. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110. https://doi.org/10.1093/bioinformatics/btac020.
20. Bordin, N., Dallago, C., Heinzinger, M., Kim, S., Littmann, M., Rauer, C., Steinegger, M., Rost, B., and Orengo, C. (2023). Novel machine learning approaches revolutionize protein knowledge. Trends Biochem. Sci. *48*, 345–359. https://doi.org/10.1016/j.tibs.2022.11.001.
21. Ferruz, N., Schmidt, S., and Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. *13*, 4348. https://doi.org/10.1038/s41467-022-32007-7.
22. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. *41*, 1099–1106. https://doi.org/10.1038/s41587-022-01618-2.
23. Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G.M., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. Nat. Biotechnol. *40*, 1617–1623. https://doi.org/10.1038/s41587-022-01432-w.
24. Shuai, R.W., Ruffolo, J.A., and Gray, J.J. (2022). Generative language modeling for antibody design. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.13.472419.
25. Olsen, T.H., Moal, I.H., and Deane, C.M. (2022). AbLang: an antibody language model for completing antibody sequences. Bioinform. Adv. *2*, vbac046. https://doi.org/10.1093/bioadv/vbac046.
26. Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., and Kim, P.S. (2024). Efficient evolution of human antibodies from general protein language models. Nat. Biotechnol. *42*, 275–283. https://doi.org/10.1038/s41587-023-01763-2.
27. Raybould, M.I.J., Kovaltsuk, A., Marks, C., and Deane, C.M. (2021). CoV-AbDab: the coronavirus antibody database. Bioinformatics *37*, 734–735. https://doi.org/10.1093/bioinformatics/btaa739.
28. Wu, N.C., and Wilson, I.A. (2017). A perspective on the structural and functional constraints for immune evasion: insights from influenza virus. J. Mol. Biol. *429*, 2694–2709. https://doi.org/10.1016/j.jmb.2017.06.015.
29. Lang, S., Xie, J., Zhu, X., Wu, N.C., Lerner, R.A., and Wilson, I.A. (2017). Antibody 27F3 broadly targets influenza A group 1 and 2 hemagglutinins through a further variation in V_H1-69 antibody orientation on the HA stem. Cell Rep. *20*, 2935–2943. https://doi.org/10.1016/j.celrep.2017.08.084.
30. Cheung, C.S.F., Fruehwirth, A., Paparoditis, P.C.G., Shen, C.H., Foglierini, M., Joyce, M.G., Leung, K., Piccoli, L., Rawi, R., Silacci-Fregni, C., et al. (2020). Identification and structure of a multidonor class of head-directed influenza-neutralizing antibodies reveal the mechanism for its recurrent elicitation. Cell Rep. *32*, 108088. https://doi.org/10.1016/j.celrep.2020.108088.
31. Wu, N.C., Yuan, M., Liu, H., Lee, C.D., Zhu, X., Bangaru, S., Torres, J.L., Caniels, T.G., Brouwer, P.J.M., van Gils, M.J., et al. (2020). An alternative binding mode of IGHV3-53 antibodies to the SARS-CoV-2 receptor binding domain. Cell Rep. *33*, 108274. https://doi.org/10.1016/j.celrep.2020.108274.
32. Yuan, M., Liu, H., Wu, N.C., Lee, C.D., Zhu, X., Zhao, F., Huang, D., Yu, W., Hua, Y., Tien, H., et al. (2020). Structural basis of a shared antibody response to SARS-CoV-2. Science *369*, 1119–1123. https://doi.org/10.1126/science.abd2321.
33. Wiley, D.C., Wilson, I.A., and Skehel, J.J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature *289*, 373–378. https://doi.org/10.1038/289373a0.
34. Caton, A.J., Brownlee, G.G., Yewdell, J.W., and Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell *31*, 417–427. https://doi.org/10.1016/0092-8674(82)90135-0.
35. Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004). Mapping the antigenic and genetic evolution of influenza virus. Science *305*, 371–376. https://doi.org/10.1126/science.1097211.

36. Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C.M., Vervaet, G., Skepner, E., Lewis, N.S., Spronken, M.I.J., Russell, C.A., et al. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science *342*, 976–979. https://doi.org/10.1126/science.1244730.

37. Wu, N.C., and Wilson, I.A. (2020). Influenza hemagglutinin structures and antibody recognition. Cold Spring Harb. Perspect. Med. *10*, a038778. https://doi.org/10.1101/cshperspect.a038778.

38. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013). GenBank. Nucleic Acids Res. *41*, D36–D42. https://doi.org/10.1093/nar/gks1195.

39. Kovaltsuk, A., Leem, J., Kelm, S., C.M. Snowden, J., Deane, C.M., and Krawczyk, K. (2018). Observed Antibody Space: a resource for data mining next-generation sequencing of antibody repertoires. J. Immunol. *201*, 2502–2509. https://doi.org/10.4049/jimmunol.1800708.

40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Preprint at arXiv.

41. Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nat. Commun. *12*, 3168. https://doi.org/10.1038/s41467-021-23303-9.

42. Nakamura, G., Chai, N., Park, S., Chiang, N., Lin, Z., Chiu, H., Fong, R., Yan, D., Kim, J., Zhang, J., et al. (2013). An in vivo human-plasmablast enrichment technique allows rapid identification of therapeutic influenza A antibodies. Cell Host Microbe *14*, 93–103. https://doi.org/10.1016/j.chom.2013.06.004.

43. Joyce, M.G., Wheatley, A.K., Thomas, P.V., Chuang, G.Y., Soto, C., Bailer, R.T., Druz, A., Georgiev, I.S., Gillespie, R.A., Kanekiyo, M., et al. (2016). Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. Cell *166*, 609–623. https://doi.org/10.1016/j.cell.2016.06.043.

44. Chen, Y., Wang, F., Yin, L., Jiang, H., Lu, X., Bi, Y., Zhang, W., Shi, Y., Burioni, R., Tong, Z., et al. (2022). Structural basis for a human broadly neutralizing influenza A hemagglutinin stem-specific antibody including H17/18 subtypes. Nat. Commun. *13*, 7603. https://doi.org/10.1038/s41467-022-35236-y.

45. Corti, D., Voss, J., Gamblin, S.J., Codoni, G., Macagno, A., Jarrossay, D., Vachieri, S.G., Pinna, D., Minola, A., Vanzetta, F., et al. (2011). A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science *333*, 850–856. https://doi.org/10.1126/science.1205669.

46. Sui, J., Hwang, W.C., Perez, S., Wei, G., Aird, D., Chen, L.M., Santelli, E., Stec, B., Cadwell, G., Ali, M., et al. (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat. Struct. Mol. Biol. *16*, 265–273. https://doi.org/10.1038/nsmb.1566.

47. Ekiert, D.C., Bhabha, G., Elsliger, M.A., Friesen, R.H.E., Jongeneelen, M., Throsby, M., Goudsmit, J., and Wilson, I.A. (2009). Antibody recognition of a highly conserved influenza virus epitope. Science *324*, 246–251. https://doi.org/10.1126/science.1171491.

48. Ekiert, D.C., Friesen, R.H.E., Bhabha, G., Kwaks, T., Jongeneelen, M., Yu, W., Ophorst, C., Cox, F., Korse, H.J.W.M., Brandenburg, B., et al. (2011). A highly conserved neutralizing epitope on group 2 influenza A viruses. Science *333*, 843–850. https://doi.org/10.1126/science.1204839.

49. Dreyfus, C., Laursen, N.S., Kwaks, T., Zuijdgeest, D., Khayat, R., Ekiert, D.C., Lee, J.H., Metlagel, Z., Bujny, M.V., Jongeneelen, M., et al. (2012). Highly conserved protective epitopes on influenza B viruses. Science *337*, 1343–1348. https://doi.org/10.1126/science.1222908.

50. Friesen, R.H.E., Lee, P.S., Stoop, E.J.M., Hoffman, R.M.B., Ekiert, D.C., Bhabha, G., Yu, W., Juraszek, J., Koudstaal, W., Jongeneelen, M., et al. (2014). A common solution to group 2 influenza virus neutralization. Proc. Natl. Acad. Sci. USA *111*, 445–450. https://doi.org/10.1073/pnas.1319058110.

51. Wu, Y., Cho, M., Shore, D., Song, M., Choi, J., Jiang, T., Deng, Y.Q., Bourgeois, M., Almli, L., Yang, H., et al. (2015). A potent broad-spectrum protective human monoclonal antibody crosslinking two haemagglutinin monomers of influenza A virus. Nat. Commun. *6*, 7708. https://doi.org/10.1038/ncomms8708.

52. Kallewaard, N.L., Corti, D., Collins, P.J., Neu, U., McAuliffe, J.M., Benjamin, E., Wachter-Rosati, L., Palmer-Hill, F.J., Yuan, A.Q., Walker, P.A., et al. (2016). Structure and function analysis of an antibody recognizing all influenza A subtypes. Cell *166*, 596–608. https://doi.org/10.1016/j.cell.2016.05.073.

53. Matsuda, K., Huang, J., Zhou, T., Sheng, Z., Kang, B.H., Ishida, E., Griesman, T., Stuccio, S., Bolkhovitinov, L., Wohlbold, T.J., et al. (2019). Prolonged evolution of the memory B cell response induced by a replicating adenovirus-influenza H5 vaccine. Sci. Immunol. *4*, eaau2710. https://doi.org/10.1126/sciimmunol.aau2710.

54. Wu, N.C., Andrews, S.F., Raab, J.E., O'Connell, S., Schramm, C.A., Ding, X., Chambers, M.J., Leung, K., Wang, L., Zhang, Y., et al. (2020). Convergent evolution in breadth of two V_H6-1-encoded influenza antibody clonotypes from a single donor. Cell Host Microbe *28*, 434–444.e4. https://doi.org/10.1016/j.chom.2020.06.003.

55. Hanke, L., Sheward, D.J., Pankow, A., Vidakovics, L.P., Karl, V., Kim, C., Urgard, E., Smith, N.L., Astorga-Wells, J., Ekström, S., et al. (2022). Multivariate mining of an alpaca immune repertoire identifies potent cross-neutralizing SARS-CoV-2 nanobodies. Sci. Adv. *8*, eabm0220. https://doi.org/10.1126/sciadv.abm0220.

56. Henry, C., Zheng, N.Y., Huang, M., Cabanov, A., Rojas, K.T., Kaur, K., Andrews, S.F., Palm, A.E., Chen, Y.Q., Li, Y., et al. (2019). Influenza virus vaccination elicits poorly adapted B cell responses in elderly individuals. Cell Host Microbe *25*, 357–366.e6. https://doi.org/10.1016/j.chom.2019.01.002.

57. Corbett, K.S., Moin, S.M., Yassine, H.M., Cagigi, A., Kanekiyo, M., Boyoglu-Barnum, S., Myers, S.I., Tsybovsky, Y., Wheatley, A.K., Schramm, C.A., et al. (2019). Design of nanoparticulate group 2 influenza virus hemagglutinin stem antigens that activate unmutated ancestor B cell receptors of broadly neutralizing antibody lineages. mBio *10*, e02810-18. https://doi.org/10.1128/mBio.02810-18.

58. Lei, R., Kim, W., Lv, H., Mou, Z., Scherm, M.J., Schmitz, A.J., Turner, J.S., Tan, T.J.C., Wang, Y., Ouyang, W.O., et al. (2023). Leveraging vaccination-induced protective antibodies to define conserved epitopes on influenza N2 neuraminidase. Immunity *56*, 2621–2634.e6. https://doi.org/10.1016/j.immuni.2023.10.005.

59. Adachi, Y., Tonouchi, K., Nithichanon, A., Kuraoka, M., Watanabe, A., Shinnakasu, R., Asanuma, H., Ainai, A., Ohmi, Y., Yamamoto, T., et al. (2019). Exposure of an occluded hemagglutinin epitope drives selection of a class of cross-protective influenza antibodies. Nat. Commun. *10*, 3883. https://doi.org/10.1038/s41467-019-11821-6.

60. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

61. Kelow, S.P., Adolf-Bryfogle, J., and Dunbrack, R.L. (2020). Hiding in plain sight: structure and sequence analysis reveals the importance of the antibody DE loop for antibody-antigen binding. mAbs *12*, 1840005. https://doi.org/10.1080/19420862.2020.1840005.

62. Andrews, S.F., Huang, Y., Kaur, K., Popova, L.I., Ho, I.Y., Pauli, N.T., Henry Dunand, C.J., Taylor, W.M., Lim, S., Huang, M., et al. (2015). Immune history profoundly affects broadly protective B cell responses to influenza. Sci. Transl. Med. *7*, 316ra192. https://doi.org/10.1126/scitranslmed.aad0522.

63. Impagliazzo, A., Milder, F., Kuipers, H., Wagner, M.V., Zhu, X., Hoffman, R.M.B., van Meersbergen, R., Huizingh, J., Wanningen, P., Verspuij, J., et al. (2015). A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. Science *349*, 1301–1306. https://doi.org/10.1126/science.aac7263.

64. Fu, Y., Zhang, Z., Sheehan, J., Avnir, Y., Ridenour, C., Sachnik, T., Sun, J., Hossain, M.J., Chen, L.M., Zhu, Q., et al. (2016). A broadly neutralizing anti-influenza antibody reveals ongoing capacity of haemagglutinin-specific memory B cells to evolve. Nat. Commun. *7*, 12780. https://doi.org/10.1038/ncomms12780.

65. Pappas, L., Foglierini, M., Piccoli, L., Kallewaard, N.L., Turrini, F., Silacci, C., Fernandez-Rodriguez, B., Agatic, G., Giacchetto-Sasselli, I., Pellicciotta, G., et al. (2014). Rapid development of broadly influenza neutralizing antibodies through redundant mutations. Nature *516*, 418–422. https://doi.org/10.1038/nature13764.

66. Benton, D.J., Nans, A., Calder, L.J., Turner, J., Neu, U., Lin, Y.P., Ketelaars, E., Kallewaard, N.L., Corti, D., Lanzavecchia, A., et al. (2018). Influenza hemagglutinin membrane anchor. Proc. Natl. Acad. Sci. USA *115*, 10112–10117. https://doi.org/10.1073/pnas.1810927115.

67. Guthmiller, J.J., Han, J., Utset, H.A., Li, L., Lan, L.Y.L., Henry, C., Stamper, C.T., McMahon, M., O'Dell, G., Fernández-Quintero, M.L., et al. (2022). Broadly neutralizing antibodies target a haemagglutinin anchor epitope. Nature *602*, 314–320. https://doi.org/10.1038/s41586-021-04356-8.

68. Andrews, S.F., Cominsky, L.Y., Shimberg, G.D., Gillespie, R.A., Gorman, J., Raab, J.E., Brand, J., Creanga, A., Gajjala, S.R., Narpala, S., et al. (2023). An influenza H1 hemagglutinin stem-only immunogen elicits a broadly cross-reactive B cell response in humans. Sci. Transl. Med. *15*, eade4976. https://doi.org/10.1126/scitranslmed.ade4976.

69. Whittle, J.R.R., Wheatley, A.K., Wu, L., Lingwood, D., Kanekiyo, M., Ma, S.S., Narpala, S.R., Yassine, H.M., Frank, G.M., Yewdell, J.W., et al. (2014). Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. J. Virol. *88*, 4047–4057. https://doi.org/10.1128/JVI.03422-13.

70. McCarthy, K.R., Lee, J., Watanabe, A., Kuraoka, M., Robinson-McCarthy, L.R., Georgiou, G., Kelsoe, G., and Harrison, S.C. (2021). A prevalent focused human antibody response to the influenza virus hemagglutinin head interface. mBio *12*, e0114421. https://doi.org/10.1128/mBio.01144-21.

71. Andrews, S.F., Joyce, M.G., Chambers, M.J., Gillespie, R.A., Kanekiyo, M., Leung, K., Yang, E.S., Tsybovsky, Y., Wheatley, A.K., Crank, M.C., et al. (2017). Preferential induction of cross-group influenza A hemagglutinin stem-specific memory B cells after H7N9 immunization in humans. Sci. Immunol. *2*, eaan2676. https://doi.org/10.1126/sciimmunol.aan2676.

72. Andrews, S.F., Chambers, M.J., Schramm, C.A., Plyler, J., Raab, J.E., Kanekiyo, M., Gillespie, R.A., Ransier, A., Darko, S., Hu, J., et al. (2019). Activation dynamics and immunoglobulin evolution of pre-existing and newly generated human memory B cell responses to influenza hemagglutinin. Immunity *51*, 398–410.e5. https://doi.org/10.1016/j.immuni.2019.06.024.

73. Teo, Q.W., Wang, Y., Lv, H., Tan, T.J.C., Lei, R., Mao, K.J., and Wu, N.C. (2023). Stringent and complex sequence constraints of an IGHV1-69 broadly neutralizing antibody to influenza HA stem. Cell Rep. *42*, 113410. https://doi.org/10.1016/j.celrep.2023.113410.

74. wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. *47*, D520–D528. https://doi.org/10.1093/nar/gky949.

75. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

76. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

77. Potocnakova, L., Bhide, M., and Pulzova, L.B. (2016). An introduction to B-cell epitope mapping and in silico epitope prediction. J. Immunol. Res. *2016*, 6760830. https://doi.org/10.1155/2016/6760830.

78. Rappazzo, C.G., Fernández-Quintero, M.L., Mayer, A., Wu, N.C., Greiff, V., and Guthmiller, J.J. (2023). Defining and studying B cell receptor and TCR interactions. J. Immunol. *211*, 311–322. https://doi.org/10.4049/jimmunol.2300136.

79. Trkola, A., Purtscher, M., Muster, T., Ballaun, C., Buchacher, A., Sullivan, N., Srinivasan, K., Sodroski, J., Moore, J.P., and Katinger, H. (1996). Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. J. Virol. *70*, 1100–1108. https://doi.org/10.1128/JVI.70.2.1100-1108.1996.

80. Lee, C.D., Watanabe, Y., Wu, N.C., Han, J., Kumar, S., Pholcharee, T., Seabright, G.E., Allen, J.D., Lin, C.W., Yang, J.R., et al. (2021). A cross-neutralizing antibody between HIV-1 and influenza virus. PLoS Pathog. *17*, e1009407. https://doi.org/10.1371/journal.ppat.1009407.

81. Mannar, D., Leopold, K., and Subramaniam, S. (2021). Glycan reactive anti-HIV-1 antibodies bind the SARS-CoV-2 spike protein but do not block viral entry. Sci. Rep. *11*, 12448. https://doi.org/10.1038/s41598-021-91746-7.

82. Bogatinovski, J., Todorovski, L., Džeroski, S., and Kocev, D. (2022). Comprehensive comparative study of multi-label classification methods. Expert Syst. Appl. *203*, 117215. https://doi.org/10.1016/j.eswa.2022.117215.

83. Yang, L., Meng, X., and Karniadakis, G.E. (2021). B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. J. Comp. Phys. *425*, 109913. https://doi.org/10.1016/j.jcp.2020.109913.

84. Schmidt, A.G., Therkelsen, M.D., Stewart, S., Kepler, T.B., Liao, H.X., Moody, M.A., Haynes, B.F., and Harrison, S.C. (2015). Viral receptor-binding site antibodies with diverse germline origins. Cell *161*, 1026–1034. https://doi.org/10.1016/j.cell.2015.04.028.

85. Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. *41*, W34–W40. https://doi.org/10.1093/nar/gkt382.

86. Soto, C., Finn, J.A., Willis, J.R., Day, S.B., Sinkovits, R.S., Jones, T., Schmitz, S., Meiler, J., Branchizio, A., and Crowe, J.E., Jr. (2020). PyIR: a scalable wrapper for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST. BMC Bioinformatics *21*, 314. https://doi.org/10.1186/s12859-020-03649-5.

87. Tareen, A., and Kinney, J.B. (2020). Logomaker: beautiful sequence logos in Python. Bioinformatics *36*, 2272–2274. https://doi.org/10.1093/bioinformatics/btz921.

88. Neumann, G., Watanabe, T., Ito, H., Watanabe, S., Goto, H., Gao, P., Hughes, M., Perez, D.R., Donis, R., Hoffmann, E., et al. (1999). Generation of influenza A viruses entirely from cloned cDNAs. Proc. Natl. Acad. Sci. USA *96*, 9345–9350. https://doi.org/10.1073/pnas.96.16.9345.

89. Wimley, W.C., and White, S.H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol. *3*, 842–848. https://doi.org/10.1038/nsb1096-842.

90. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

91. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

92. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.11692.

93. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

94. Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Software *6*, 3021. https://doi.org/10.21105/joss.03021.

95. Guthmiller, J.J., Dugan, H.L., Neu, K.E., Lan, L.Y.L., and Wilson, P.C. (2019). An efficient method to generate monoclonal antibodies from human B cells. Methods Mol. Biol. *1904*, 109–145. https://doi.org/10.1007/978-1-4939-8958-4_5.

96. Wu, N.C., Grande, G., Turner, H.L., Ward, A.B., Xie, J., Lerner, R.A., and Wilson, I.A. (2017). In vitro evolution of an influenza broadly neutralizing antibody is modulated by hemagglutinin receptor specificity. Nat. Commun. *8*, 15371. https://doi.org/10.1038/ncomms15371.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
6x-His Tag Monoclonal Antibody (HIS.H8)	Thermo Fisher Scientific	Cat# 14-6657-82; RRID:AB_2572898
HRP Rat Anti-Mouse Ig, k Light Chain	BD Biosciences	Cat# 559751; RRID:AB_397315
Goat anti-Human IgG (H+L) Secondary Antibody, HRP	Thermo Fisher Scientific	Cat# A18805; RRID:AB_2535582
Bacterial and virus strains		
NEB 5-alpha Competent E. coli	New England Biolabs	Cat# C2987H
A/Puerto Rico/8/1934 (H1N1)	N/A	N/A
A/Michigan/45/2015 (H1N1)	N/A	N/A
Chemicals, peptides, and recombinant proteins		
Lipofectamine 2000	Fisher Scientific	Cat# 11668-019
Cellfectin II Reagent	Gibco	Cat# 10362-100
TPCK-Trypsin	Thermo Fisher Scientific	Cat# 20233
Tween 20	Fisher Scientific	Cat# BP337-100
H3N2 A/Darwin/9/2021	BEI Resources	cat #: NR-59305
H3N8 A/duck/Shantou/1283/2001	BEI Resources	cat #: NR-28916
H4N6 A/mallard/Alberta/455/2015	BEI Resources	cat #: NR-51128
H7N3 A/Canada/rv444/2004	BEI Resources	cat #: NR-43740
H7N9 A/Hong Kong/125/2017	BEI Resources	cat #: NR-51367
H7N9 A/Guangdong/17SF003/2016	BEI Resources	cat #: NR-51203
H7N9 A/Hunan/02285/2017	BEI Resources	cat #: NR-51195
H7N9 A/Anhui/1/2013	BEI Resources	cat #: NR-44081
H7N9 A/Shanghai/1/2013	BEI Resources	cat #: NR-44079
H10N8 A/Jiangxi-Donghu/346/2013	BEI Resources	cat #: NR-49440
1-Step Ultra TMB-ELISA Substrate Solution	Thermo Fisher Scientific	Cat# 34028
Critical commercial assays		
QIAprep Spin Miniprep Kit	Qiagen	Cat# 27106
ZymoPure Midiprep (50 preps)	Fisher	Cat# NC0919795
NEBuilder HiFi DNA Assembly Master Mix	New England Biolabs	Cat# E2621L
Deposited data		
Collection of antibody information	This study	Table S1
Cryo-EM map of 310-18A5 Fab + SI06 HA	This study	EMDB: EMD-41849
Custom scripts and model	This study	https://doi.org/10.5281/zenodo.11359149
Experimental models: Cell lines		
Sf9 cell	ATCC	CRL-1711; PRID:CVCL_0549
MDCK-SIAT1 cells	Sigma-Aldrich	Cat# 05071502-1VL
HEK293T cells	N/A	N/A
Expi293F Cells	Thermo Fisher Scientific	Cat# A14527
Recombinant DNA		
pFastBac-miniHA-H1	This study	N/A
pFastBac-miniHA-H3	This study	N/A
phCMV3-Ab IgG heavy chain	This study	N/A
phCMV3-Ab Fab heavy chain	This study	N/A
phCMV3-Ab IgG kappa light chain	This study	N/A
phCMV3-Ab IgG lambda light chain	This study	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Octet analysis software 9.0	Sartorius	N/A
R	https://www.r-project.org/	N/A
Python	https://www.python.org/	N/A
IgBLAST	Ye et al.[85]	N/A
PyIR	Soto et al.[86]	N/A
Logomaker	Tareen and Kinney[87]	N/A
Pytorch	https://pytorch.org/	N/A
Transformers	https://huggingface.co/docs/transformers/en/index	N/A
Other		
Octet Anti-Penta-HIS (HIS1K) Biosensors	Sartorius	Cat# 18-5120
Nunc MaxiSorp flat-bottom 96 well plate	Thermo Fisher Scientific	Cat# 44-2404-21
Microplate, 96 Well, PP, F-Bottom	Grenier	Cat# 655209
Sf-900 II SFM	Thermo Fisher Scientific	Cat# 10902088
MEM medium	Thermo Fisher Scientific	Cat# 11095098
DMEM medium	Thermo Fisher Scientific	Cat# 11995065
Opti-MEM I Reduced Serum Medium	Thermo Fisher Scientific	Cat# 31985070
GlutaMAX Supplement	Thermo Fisher Scientific	Cat# 35050061
Trypsin-EDTA (0.25%), phenol red	Thermo Fisher Scientific	Cat# 25200056
Penicillin-Streptomycin	Thermo Fisher Scientific	Cat# 15140122
Fetal Bovine Serum (FBS)	Thermo Fisher Scientific	Cat# 16000044
Phosphate-buffered saline (PBS), 1X	VWR	Cat# 21-040-CM

RESOURCE AVAILABILITY

Lead contact

Information and requests for resources should be directed to and will be fulfilled by the lead contact, Nicholas C. Wu (nicwu@illinois.edu).

Materials availability

All plasmids generated in this study are available from the lead contact without restriction.

Data and code availability

- The curated influenza antibody dataset is in Table S1.
- The cryoEM map of 310-18A5 Fab in complex with SI06 HA has been deposited in the Electron Microscopy Data Bank (EMDB) with accession code EMD-41849.
- Custom python scripts for all analyses and model training have been deposited to: http://www.doi.org/10.5281/zenodo.11359137.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODELS AND STUDY PARTICIPANTS DETAILS

Cell lines

HEK293T cells (human embryonic kidney cells, female) and MDCK-SIAT1 cells (Madin-Darby canine kidney with overexpression of human 2,6-sialtransferase, female, Sigma-Aldrich) were cultured in Dulbecco's modified Eagle's medium (DMEM high glucose; Gibco) supplemented with 10% heat-inactivated fetal bovine serum (FBS; Gibco), 1% penicillin-streptomycin (Gibco), and 1× GlutaMax (Gibco). Cell passaging was performed every 3 to 4 days using 0.05% Trypsin-EDTA solution (Gibco). Expi293F cells (human embryonic kidney cells, female, ATCC) were maintained in Expi293 Expression Medium (Thermo Fisher Scientific). Sf9 cells (*Spodoptera frugiperda* ovarian cells, female, ATCC) were maintained in Sf-900 II SFM medium (Thermo Fisher Scientific).

Recombinant influenza virus

Recombinant influenza viruses were rescued using the eight-plasmid reverse genetics system.[88] Briefly, plasmids encoding the eight segments of the influenza genome were transfected into a co-culture of HEK293T cells and MDCK-SIAT1 cells (ratio of 6:1) at 60%. Transfection was performed using Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions. At 24 hours post-transfection, cells were washed twice with PBS and cell culture medium was replaced with OPTI-MEM medium supplemented with 0.8 μg mL^{-1} Tosyl phenylalanyl chloromethyl ketone (TPCK)-trypsin. The virus in the supernatant was harvested at 48 hours post-transfection and plaque-purified before growing to a high titer stock in MDCK-SIAT1 cells. All plaque-purified viruses were sequenced to confirm the absence of mutations.

METHOD DETAILS

Collections of antibody information

Sequences of each human monoclonal antibody were from the original papers and/or NCBI GenBank database (Tables S1 and S3).[38] For influenza HA antibodies, additional information, including binding specificity, donor IDs and PDB codes, was collected from the original papers (Table S1). Putative germline genes were identified by IgBLAST.[85,86] Some studies isolated antibodies from multiple donors, but the donor identity for each antibody was not always clear. For example, some studies mixed B cells from multiple donors before isolating individual B cell clones. Since the donor identity could not be distinguished among those antibodies, we considered them from the same donor with "donors", "vaccinees", "patients", or "cohorts" as the suffix of the donor ID. In addition, although two studies by Andrews et al.[71,72] had shared donors from the same clinical trial (VRC 315, ClincialTrials.gov identifier NCT02206464), their antibody naming schemes were different. The IDs for these donors had a prefix "315" as described in the first study.[71] While the prefixes of antibody names from the first study matched the donor ID (e.g. antibody 315-02-1F07 was from donor 315-02),[71] some antibody names from the second study did not (e.g. antibody name with prefix "20A-518-30").[72] Therefore, we assigned the donor ID to the antibodies from the second study by CDR H3 clustering. For example, since all CDR H3 clusters that contained antibodies with prefix 20A-605-30 also contained antibodies from 315-02, antibodies with prefix 20A-605-30 were assigned with a donor ID of 315-02.

Identification of clonotypes and public clonotypes

Using a deterministic clustering approach, CDR H3 sequences that had the same length and at least 80% amino acid sequence identity were assigned to the same CDR H3 cluster. As a result, CDR H3 of every antibody in a CDR H3 cluster would have >20% difference in amino acid sequence identity with that of every antibody in another CDR H3 cluster. A clonotype was defined as antibodies that shared the same IGHV/IGK(L)V genes with CDR H3s from the same CDR H3 cluster. A public clonotype was defined as a clonotype with antibodies from at least two donors. The epitope of each public clonotype was defined by its members. None of the public clonotypes contained antibodies targeting different epitopes.

Germline gene usage analysis

To avoid being confounded by B-cell clonal expansion, a single clonotype from the same donor was considered as one antibody that represented the consensus sequence of the given clonotype. While all antibodies within a clonotype had the same IGHV/IGK(L)V genes (see above), they may not have the same IGHD gene, often due to ambiguity in IGHD-gene assignment by IgBlast. For germline gene usage analysis, the most common IGHD gene within a clonotype from the same donor was considered.

Hydrophobic score of CDR H3

The hydrophobic score for a CDR H3 with a length n was computed as follow:

$$Hydrophobic\ score = -10 \times \frac{\sum_{i=1}^{n} WW(amino\ acid_i)}{n}$$

where WW represents the Wimley-White whole residue hydrophobicity scale[89] and $amino\ acid_i$ represents the amino acid at position i. A higher hydrophobic score represents higher hydrophobicity. If the CDR H3 had an odd number of residues, the CDR H3 tip was defined as the three residues at the center of the CDR H3 sequence. If the CDR H3 had an even number of residues, the CDR H3 tip was defined as the four residues at the center of the CDR H3 sequence. The hydrophobic score of CDR H3 tip was computed in the same manner as that of CDR H3. To avoid being confounded by B-cell clonal expansion, a single clonotype from the same donor is considered as one antibody, in which the CDR H3 sequence represented the consensus among all members in the given clonotype.

Datasets for model pre-training

A total of 267,871 paired antibody sequences from memory B cell sequencing data were downloaded from Observed Antibody Space database (BType = Memory-B-Cells).[39] In addition, 12,487 paired antibody sequences were downloaded from NCBI GenBank database.[38] These antibody sequences were compiled into a single dataset and deduplicated by 95% sequence identity threshold. The deduplicated dataset was then partitioned into training (n = 229,773), validation (n = 15,375) and test sets (n = 8,660). The test set was generated by random sampling with different levels of maximum sequence identity to the training set (50%, 60%, 70%, 80%, and 90%), allowing robust evaluation of model performance. Of note, 90% maximum sequence identity indicated that

none of the antibody sequences in the test set had >90% sequence identity with any of the sequences in the training set. In other words, the highest pairwise sequence identity between the test and training sets was 90%. To generate a balanced and robust training set, we implemented an upsampling technique based on the IGK(L)V genes. Specifically, we identified IGK(L)V genes with less than 5,000 counts and then performed random sampling to augment the dataset, ensuring each of these IGK(L)V genes had precisely 5,000 sequences. After upsampling, our training set had 467,018 paired antibody sequences. Upsampling only applied to the training set, but not the validation and test sets.

To evaluate the performance of our model, perplexity was measured using the test set. Perplexity was defined as the exponential of the negative log-likelihood of the sequence as follows:

$$Perplexity(x) = \exp\left\{-\log p\left(x_{i\epsilon M}\middle|x_{j\notin M}\right)\right\}$$

where mask M is a random variable denoting a set of tokens from input sequence x. Specifically, 15% of the input sequence underwent replacement by the mask M. Here, the perplexity was calculated as exponent of the loss obtained from the model. This approach provided a measure of how well the model aligned with the original sequence, where lower perplexity values indicated a better fit. The perplexity of our mBLM was 2.63, whereas that of ESM2 is 3.62.

Sequences of antibodies with known specificities for model fine-tuning

Sequences of antibodies to "HA:Head" (influenza HA head) and "HA:Stem" (influenza HA stem) were from the curated dataset in this present study. Sequences of antibodies to "S:NTD" (SARS-CoV-2 spike NTD), "S:RBD" (SARS-CoV-2 spike RBD), and "S:S2" (SARS-CoV-2 spike S2) were from our previous study.[15] Sequences of antibodies to "HIV" (human immunodeficiency virus) and "Others" (none of the above) were collected from NCBI GenBank database.[38] Antibodies to "HIV" were classified as those from GenBank with the word "HIV" in the "References" or "Description" fields. Here, only heavy chain variable domain sequences were used for model fine-tuning. We performed sequence clustering with varying sequence identity cutoff (50%, 60%, 70%, 80%, 90%, and 95%) using cd-hit (-M 32000 -d 0 -T 8 -n 5 -aL 0.8 -s 0.95 -uS 0.2 -sc 1 -sf 1).[90] We observed that at a cutoff of 90% sequence identity, sequences of antibodies with different specificities could be found within the same cluster, indicating that a stringent sequence identity cutoff of >90% was needed for accurate specificity prediction by traditional sequence clustering method. Based on this result, antibodies with unknown specificities, but shared >90% sequence identity with any antibody that belonged to "HA:Head", "HA:Stem", "HIV", "S:NTD", "S:RBD", or "S:S2", were discarded and not assigned to the "Others" category. Our final dataset for model fine-tuning contained the heavy chain sequences from a total of 388 antibodies to "HA:Head", 509 antibodies to "HA:Stem", 6,995 antibodies to "HIV", 399 antibodies to "S:NTD", 4112 antibodies to "S:RBD", 682 antibodies to "S:S2", and 15,043 antibodies to "Others". This dataset was then partitioned into training, validation and test sets, with an approximate ratio of 8:1:1. To split the sequences between training and test sets, all sequences in the training and test sets were first clustered using cd-hit using a sequence identity cutoff of 80%.[90] Then sequences within the same cluster would either be assigned to the training set or test set. In other words, sequences in the training set and sequences in the test set would not fall into the same cluster. We further applied a manual cleanup to ensure that the training and test sets had a minimum pairwise Levenshtein distance of 10. We also applied the upsampling technique to the training set to ensure the number of antibody sequences in different specificity categories was balanced.

Pre-trained memory B cell language model (mBLM)

Masked Language Modeling (MLM)

Masked language modeling such as Bidirectional Encoder Representations from Transformers (BERT)[91] has been shown as a powerful pretraining technique for language models, enabling contextual information to be captured and generalized to various downstream tasks. Here, mBLM was trained to predict the masked amino acids of input sequence based on surrounding context:

$$\mathcal{L}_{MLM} = -\sum_{i\epsilon M}\log p\left(x_i\middle|x_{context}\right)$$

where M represents a randomly generated mask that includes 15% of positions i in the sequence x_i. The model was tasked with predicting the identity of the amino acids x_i in the mask from the surrounding context $x_{context}$. Being trained to predict masked tokens, mBLM learned to understand the relationships between amino acid residues in a sequence, leading to a robust and effective language representations.

mBLM architecture

We adapted RoBERTa[92] as the basic model architecture, with the following hyperparameters:

Tokenizer: ESM2[18]
Token length: 150
Number of Layers: 6
Number of Attention heads: 12
Embedding dimension: 768
Feed-Forward Hidden Size: 3072
Dropout: 0.1

mBLM pre-training

mBLM was pre-trained with a context size of 250 tokens, which represented the amino acid sequences of both heavy and light chain variable domains. Since the total length of heavy chain and light chain variable domains was generally less than 250 amino acids, separation tokens were added in between. We adapted tokenizer from ESM2,[18] which converted amino acids into numerical representations (a total of 33 tokens including special tokens like [MASK]). The model was trained by masked language modeling (MLM) as described above. The model was optimized using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, ϵ = 1e-08, and a learning rate of 5e-05. The model was trained using Huggingface transformers toolkit and efficiently distributed across one NVIDIA A100 and three NVIDIA RTX A5000. The entire pre-training process was completed within 24 hours, showcasing the efficiency and scalability of our approach. Different model architectures were explored for the pre-training process. In general, increasing the model depth (in terms of number of layers) improved performance when the number of layers was small, yet no significant improvement was observed beyond six layers.

Model fine-tuning for specificity prediction

Model details

The final-layer embeddings from the pre-trained mBLM were extracted as the initial hidden state for the specificity prediction model. This initial state was then fed through a multi-head self-attention block and a multi-layer perceptron (MLP) block. An attention block was incorporated between the mBLM embeddings and the MLP significantly to enhance model interpretability. Within the attention block, the self-attention layer was followed by a layer normalization to normalize the output. Subsequently, an adaptive average pooling was applied to the attended representation to aggregate information across sequence dimension, resulting in a fixed size tensor with a shape that was defined by batch size and hidden dimension. The flattened tensor was then passed through the MLP block, comprising a series of fully connected layers, ReLU activation functions, and dropout operations. These layers transformed the high-dimensional representation to low-dimensional features. Finally, the output was passed through a fully connected layer with seven output units, each represented one of the seven specificity categories. Of note, experimenting with different layers of classifier did not yield significant difference in performance.

Resampling procedure

To assess the robustness of our mBLM in predicting antibody specificity, a tailored resampling technique was employed. This involved random down/upsampling of the curated dataset for each specificity category, ensuring a balanced representation during model training. Subsequently, the dataset was randomly split based on sequence similarity as described above. These resampling and splitting processes were iterated 15 times, generating different training and test sets in each round. Subsequently, the model underwent 15 rounds of training and testing, and its performance was evaluated for each round. The overall model performance was quantified as the average across all 15 rounds.

mBLM fine-tuning

The model was trained using the PyTorch Lightning framework using Adam optimizer with a learning rate of 2e-05 and a batch size of 32. Early stopping was applied to monitor the validation loss.

ESM2 fine-tuning

Similar to mBLM fine-tuning, the final representations of ESM2 model (33 layers and 650 million parameters) were extracted as the initial hidden state for specificity prediction. This initial state was then fed through the attention and MLP blocks. The model was trained using the PyTorch Lightning framework using Adam optimizer with a learning rate of 1e-04 and a batch size 32. Early stopping was applied to monitor the validation loss. The best model checkpoint was saved.

kNN classifier

The k-Nearest Neighbors (KNN) classification algorithm was implemented to predict antibody specificity using the sklearn.neighbors.KNeighborsClassifier module.[93] Briefly, the antibody sequences were encoded via the one-hot encoding scheme, and the encoded sequences were fed into the KNN classifier, with a range of k values, including 1, 3, 5, 10, 20, 30, 50, 100, and 500. F1 score and confusion matrices were computed to evaluate classification accuracy for each k value. The optimal k value was selected based on the resulting confusion matrices.

Performance Metrics

The fine-tuned model was evaluated using the micro F1 score, which counts the total true positives, false negatives and false positives and represents the globally arithmetic mean of the harmonic means of precision and recall, as well as confusion matrix. The calculations were conducted using sklearn metrics functions using a threshold of 0.5 for class labeling.[93]

Model Interpretation

Gradient-weighted Class Activation Mapping (Grad-CAM) analysis

Grad-CAM, which is a class-discriminative localization technique that provides visual explanations for predictions made by CNN-based models,[40] was used to identify residues in a protein sequence that are important for the prediction of a particular function.[41] To calculate Grad-CAM, we first computed the importance weights α_i^c for the input sequence:

$$\alpha_i^c = \frac{1}{D}\sum_{d \in D}\frac{\partial y^c}{\partial x_d^i}$$

where α_i^c represents the global average pooling over embedding dimension D for the importance weights of residue i for predicting specificity class c. Then, the saliency map was obtained in a residue space by generating the weighted forward activation maps A^i, followed by:

$$S_i^c = max(0, \alpha_i^c A^i)$$

where S_i^c represents the relative importance (saliency score) of residue i to specificity class c. This function ensured that only features with positive influence on the functional label were preserved.

Saliency map clustering

We applied hierarchical clustering with Ward's method to perform saliency map clustering. Euclidean distance was used to calculate the distance matrix that quantified the pairwise dissimilarity between saliency maps. We then used the linkage function to define the hierarchical relationships between the samples. Finally, clustered results were visualized using clustermap function in seaborn.[94] Saliency map indicated different "attentions" on the final-layer embeddings of pre-trained mBLM. Therefore, clustering could also be performed using the final layer embeddings of the pre-trained mBLM. However, clustering on final layer embeddings will result in clusters whose saliency maps were too divergent to be amenable to downstream sequence motif analysis.

Sequence logo analysis

To identify antibody sequence features within each cluster, we employed a thresholding approach based on the saliency scores. Specifically, for each antibody sequence, residues with a saliency score >0.5 were identified. The amino acid sequences at those residues were then extracted. For each residue position in each cluster, extracted amino acid sequences with the same identity were counted. Sequence logos were generated by Logomaker in Python,[87] with the height of each amino acid proportion to its count.

Structural analysis of saliency score

For those HA stem antibodies with structural information available, the relationship between saliency score of each residue and its minimum distance to HA was examined. Distance was calculated using the application programming interface in PyMOL (Schrödinger).

Expression and purification of mini-HA and HA

The H1 mini-HA (#4900),[63] H3 mini-HA,[57] H1N1 A/Solomon Island/3/2006 HA were fused with N-terminal gp67 signal peptide and a C-terminal BirA biotinylation site, thrombin cleavage site, trimerization domain, and a 6xHis-tag, and then cloned into a customized baculovirus transfer vector.[48] Subsequently, recombinant bacmid DNA was generated using the Bac-to-Bac system (Thermo Fisher Scientific) according to the manufacturer's instructions. Baculovirus was generated by transfecting the purified bacmid DNA into adherent Sf9 cells using Cellfectin reagent (Thermo Fisher Scientific) according to the manufacturer's instructions. The baculovirus was further amplified by passaging in adherent Sf9 cells at a multiplicity of infection (MOI) of 1. Recombinant mini-HA protein was expressed by infecting 1 L of suspension Sf9 cells at an MOI of 1. On day 3 post-infection, Sf9 cells were pelleted by centrifugation at 4000 × *g* for 25 min, and soluble recombinant mini-HA and HA were purified from the supernatant by affinity chromatography using Ni Sepharose excel resin (Cytiva) and then size exclusion chromatography using a HiLoad 16/100 Superdex 200 prep grade column (Cytiva) in 20 mM Tris-HCl pH 8.0, 100 mM NaCl. The purified mini-HA protein was concentrated by Amicon spin filter (Millipore Sigma) and filtered by 0.22 μm centrifuge tube filters (Costar). Concentration of the protein was determined by nanodrop (Fisher Scientific). Proteins were subsequent aliquoted, flash frozen by dry-ice ethanol mixture, and stored at -80 °C until used. HA proteins from the following strains were obtained from BEI Resources: H3N2 A/Darwin/9/2021 (cat #: NR-59305), H3N8 A/duck/Shantou/1283/2001 (cat #: NR-28916), H4N6 A/mallard/Alberta/455/2015 (cat #: NR-51128), H7N3 A/Canada/rv444/2004 (cat #: NR-43740), H7N9 A/Hong Kong/125/2017 (cat #: NR-51367), H7N9 A/Guangdong/17SF003/2016 (cat #: NR-51203), H7N9 A/Hunan/02285/2017 (cat #: NR-51195), H7N9 A/Anhui/1/2013 (cat #: NR-44081), H7N9 A/Shanghai/1/2013 (NR-44079), and H10N8 A/Jiangxi-Donghu/346/2013 (cat #: NR-49440).

Expression and purification of IgG

The heavy and light chain genes of the obtained antibody were synthesized as eBlocks (Integrated DNA Technologies), and then cloned into human IgG1 and human kappa or lambda light chain expression vectors using Gibson assembly according to a previously described method.[95] The plasmids were transiently co-transfected into HEK293T cells at a mass ratio of 2:1 (HC:LC) using Lipofectamine 2000 (Thermo Fisher Scientific). On day 3 post-transfection, supernatant containing the IgG was collected for binding experiment. The expression of IgG was confirmed by SDS-PAGE gel electrophoresis and Coomassie Blue R-250 staining. Selected IgGs were purified using a CaptureSelect CH1-XL Pre-packed Column (Thermo Fisher Scientific).

Expression and purification of Fab

Fab heavy and light chains were cloned into phCMV3 vector. The plasmids were transiently co-transfected into Expi293F cells at a mass ratio of 2:1 (HC:LC) using ExpiFectamine 293 Reagent (Thermo Fisher Scientific). After transfection, the cell culture supernatant was collected at 6 days post-transfection. The Fab was then purified using a CaptureSelect CH1-XL pre-packed column (Thermo Fisher Scientific).

Enzyme-linked immunosorbent assay (ELISA)

Nunc MaxiSorp ELISA plates (Thermo Fisher Scientific) were utilized and coated with 100 μL of recombinant proteins at a concentration of 1 μg ml^{-1} in a 1× PBS solution. The coating process was performed overnight at 4°C. On the following day, the ELISA plates were washed three times with 1× PBS supplemented with 0.05% Tween 20, and then blocked using 200 μL of 1× PBS with 5% non-fat milk powder for 2 hours at room temperature. After the blocking step, 100 μL of IgGs from the supernatant were added to each well and incubated for 2 hours at 37°C. The ELISA plates were washed three times to remove any unbound IgGs. Next, the ELISA plates were incubated with horseradish peroxidase (HRP)-conjugated goat anti-human IgG antibody (1:5000, Invitrogen) for 1 hour at 37°C. Subsequently, the ELISA plates were washed five times using PBS containing 0.05% Tween 20. Then, 100 μL of 1-Step Ultra TMB-ELISA Substrate Solution (Thermo Fisher Scientific) was added to each well. After 15 min incubation, 50 μL of 2 M H_2SO_4 solution was added to each well. The absorbance of each well was measured at a wavelength of 450 nm using a Sunrise absorbance microplate reader (BioTek Synergy HTX Multimode Reader).

Biolayer interferometry binding assay

Binding assays were performed by biolayer interferometry (BLI) using an Octet Red96e instrument (FortéBio) at room temperature as described previously.[96] Briefly, His-tagged mini-HA proteins at 0.5 μM in 1× kinetics buffer (1× PBS, pH 7.4, 0.01% w/v BSA and 0.002% v/v Tween 20) were loaded onto anti-Penta-HIS (HIS1K) biosensors and incubated with the indicated concentrations of Fab or IgG. The assay consisted of five steps: (1) baseline: 60 s with 1× kinetics buffer; (2) loading: 60 s with His-tagged mini-HA; (3) baseline: 60 s with 1× kinetics buffer; (4) association: 60 s with Fab or IgG samples; and (5) dissociation: 60 s with 1× kinetics buffer. For estimating the exact K_D, a 1:1 binding model was used.

Virus neutralization assay

MDCK-SIAT1 cells were seeded in a 96-well, flat-bottom cell culture plate (Thermo Fisher). The next day, serially diluted monoclonal antibodies were mixed with an equal volume of virus and incubated at 37°C for 1 hour. The antibody/virus mixture was then incubated with the MDCK-SIAT1 cells at 37°C after the cells were washed twice with PBS. Following a 1-hour incubation, the antibody/virus mixture was replaced with Minimum Essential Medium (MEM) supplemented with 25 mM of 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) and 1 μg mL^{-1} of TPCK-trypsin. The plate was incubated at 37°C for 72 hours and the presence of virus was detected by hemagglutination assay. The results were analyzed using Prism software (GraphPad).

Cryogenic electron microscopy (cryo-EM) analysis

To prepare cryoEM grid, an aliquot of 4 μL purified protein at ~0.5 mg mL^{-1} concentration with 7.5 μM lauryl maltose neopentyl glycol (LMNG) was applied to a 200-mesh Quantifoil 2Um Cu grid that was pre-treated with glow-discharge. Subsequently, the grid was blotted in a Vitrobot Mark IV machine (force = 0, time = 3 seconds), and plunge-frozen in liquid ethane. The grid was then loaded in a ThermoFisher Glacios microscope with a Volta Phase Plate and Falcon4 Direct Electron Detector. Data collection was done with Smart EPU software. Images were recorded at 130,000× magnification, corresponding to a pixel size of 0.96 Å/pix at super-resolution mode of the camera. A defocus range of -0.6 μm to -3 μm was set. A total dose of 52.76 $e^{-}/Å^2$ of each exposure was fractionated into 40 frames. CryoEM data processing was performed with cryoSPARC v4.3.0 following regular single-particle procedures. The CryoEM experiment was performed at the UIUC Materials Research Laboratory Central Research Facilities. Statistics are provided in Table S4. Structure was visualized using UCSF ChimeraX v1.5 (UCSF).

QUANTIFICATION AND STATISTICAL ANALYSIS

Standard deviation for K_D estimation was computed by Octet analysis software 9.0. Student's t-tests were performed in R.