bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

[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

* These authors contributed equally to this work

§ To whom correspondence may be addressed. Email: nicwu@illinois.edu (N.C.W.)

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

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 inaccessibility of datasets for model training. In this study, we curated a dataset of >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 captured key sequence motifs 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 antibody response to influenza virus, but also provides an invaluable resource for applying deep learning to antibody research.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Discovery and characterization of monoclonal antibodies are central to the understanding of human immune response, as well as design of vaccines and therapeutics [1, 2]. As exemplified by 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 cryogenic electron 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.

Recently, the application of natural language processing has revolutionized protein structure and function prediction as well as protein design [17-23]. While several language models for antibodies have also been developed [24-26], none of them enables antibody specificity prediction to the best of our knowledge. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

many studies have reported sequences of antibodies with known epitopes, such information is often not centralized. Database such as CoV-AbDab, which documents the sequence and epitope information for >10,000 antibodies to coronavirus [27], is 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 conserved stem domain [28]. In this study, we manually curated 5,561 human antibodies to influenza hemagglutinin (HA) protein from research publications and patents. Recurring sequence features among these HA antibodies were identified, many of which were previously unknown. 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

A large-scale collection of influenza antibody information

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, 564 (10.1%) bind to the globular head domain and 518 (9.3%) bind to the stem domain. Epitope information was not available for the remaining 4,479 HA antibodies.

HA head and stem antibodies have distinct sequence features

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure 1A, Figure 1C, and Figure 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 (**Figure S2 and 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 S3B, Figure S4A, and Table S2**). Most of these antibodies contain an IGHD4-17-encoded YGD motif in the complementarity determining region (CDR) H3 (**Table S2**). Consistently, 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 S4B and 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 SARS-CoV-2 spike (S) receptor-binding domain (RBD) [31, 32].

While the major antigenic sites in the HA head domain largely consist of hydrophilic and charged amino acids [33-36], HA stem antibodies are known to commonly target a hydrophobic groove [37]. Consistently, the CDR H3 sequences of HA stem antibodies had significantly higher

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

hydrophobicity than those of HA head antibodies (p = 0.001) (**Figure 2A**). Such 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 = 4e-12) (**Figure 2B**). In contrast, the CDR H3 lengths of antibodies to HA head and stem domains did not differ significantly (p = 0.38) (**Figure 2C**). Overall, these analyses reveal distinct recurring sequence features between HA head and stem antibodies.

Antibody specificity prediction using mBLM

Our previous work has shown that antibodies with different specificities can be distinguished using a sequence-based machine learning model that has a simple architecture with one transformer encoder for each CDR, followed by a multi-layered perceptron ("CDR encoders") [15]. Here, we postulated that a language model could offer better performance, given the recent success of applying language models to predict protein structures and functions [17-23]. Specifically, we aimed to pre-train a memory B cell language model (mBLM) to learn the intrinsic "grammar" of functional antibodies, and to subsequently distinguished 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 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). Since many antibodies in these specificity categories did not have light chain sequence available, only heavy chain sequences were used for specificity prediction

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(**see Methods**). Of note, the highest pairwise sequence identity between the test and training sets was 80%. In other words, the pairwise sequence identity between individual antibody sequences in the test set and the training set was at least 20% (i.e. 26 amino acids). As indicated by the confusion matrix analysis and F1 score, mBLM had a decent performance on the test set (**Figure 3B-C**). The F1 score on the test set was 0.75 for mBLM, but only 0.49 for CDR encoders (**Figure 3C**). The performance of mBLM, which had 41 million parameters, was also slightly better than the pre-trained general protein language model ESM2 with 650 million parameters (F1 score on the test set = 0.74) [18]. This result demonstrates that mBLM is an efficient 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. Recent advancements in the field of computer vision have employed Gradient-Weighted Class Activation Maps (Grad-CAMs) on 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 acid residues for antibody specificity prediction. Our result indicates that residues with high importance, as indicated by the saliency score, were enriched in the CDRs (**Figure 4A**).

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 are 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure S5A-B**) [42], IGHV1-69 with Y98 (**Figure S5A-D**) [8], and IGHV6-1 with an FGV motif in the CDR H3 (**Figure S5E-F**) [43].

When we projected the saliency score of individual residues on the structures, residues closer to the epitope appeared to have a higher saliency score (**Figure 4D and Figure S5G-I**). Consistently, through systematically analyzing 18 structures of HA stem antibodies [7, 29, 42, 44-54], 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**). Together, our result indicates that the fine-tuned mBLM could identify residues that were critical for binding and utilized them for specificity prediction, despite structural information was not used for model training.

To gain additional insights into the learned features of mBLM, we analyzed the final-layer embeddings of the pre-trained mBLM using t-SNE (t-distributed Stochastic Neighbor Embedding). 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 S6A**). Moreover, antibodies from the same specificity category also tended to cluster together (**Figure S6B**). These observations demonstrated that even during the pre-training step, mBLM partially learned the sequence features that were determinants for antibody specificity, hence specificity prediction.

Discovering HA stem antibodies using mBLM

There are two non-overlapping epitopes in the HA stem, namely central stem epitope [44, 45] and anchor stem epitope [55, 56]. A recent study has reported the isolation of 60 HA antibodies to the central stem epitope, and 38 to the anchor stem epitope [57]. While these antibodies were not in

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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, the fine-tuned mBLM correctly predicted 67% (40/60) as HA stem antibodies (**Figure 5A**). In contrast, among the 38 antibodies to the anchor stem epitope, only 8% (3/38) were predicted as HA stem antibodies (**Figure 5A**). 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 two years ago [56]. 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% (4,479/5,561) have unknown epitopes, of which 4,452 have heavy chain sequence information available. Subsequently, we applied the fine-tuned mBLM to predict the specificities of these 4,452 antibodies. While 40% (1,769/4,452) were predicted as HA stem antibodies, only 3% (119/4,452) were predicted as HA head antibodies (**Figure 5B**). HA head antibodies were expected to have a much higher sequence diversity than HA stem antibodies, because the HA head domain has a huge sequence diversity across influenza strains and subtypes, unlike the highly conserved HA stem domain [28]. Consequently, the poor performance of the fine-tuned mBLM on HA head antibodies was likely due to insufficient sequences of HA head antibodies in our training set.

To experimentally validate our prediction result, 18 antibodies that were predicted to target HA stem were individually expressed and tested for binding to mini-HA, which is an HA stem-based construct without the HA head domain [58]. Our enzyme-linked immunosorbent assay (ELISA) result showed that 83% (15/18) could bind to mini-HA (**Figure 5C**). The remaining 3 antibodies also exhibited binding to mini-HA when tested at a high concentration (**Figure S7A**). We further selected one of the validated HA stem antibodies, 310-18A5, for additional characterization.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 S7B**) as well as mini-HA (K_D = 1.0 nM, **Figure S7C**). Besides, 310-18A5 had neutralization activity against two antigenically distinct H1N1 strains (**Figure S7D**). Consistently, cryo-EM analysis confirmed that 310-18A5 bound to the HA stem domain (**Figure 5D-E and Table S4**). Overall, these results demonstrate that the fine-tuned mBLM enables discovery of antibodies to known epitopes.

DISCUSSION

While influenza HA antibodies have been studied over decades, there has been a lack of effort to summarize the 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 have previously been reported in individual studies [7, 8, 29, 30, 42, 43, 56], 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 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 efforts 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. Although our work here applied mBLM to predict antibody specificity based on seven specificity categories, it can be fine-

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tuned to extend to any specificities as long as sufficient and diverse antibody sequences with such specificities are available. Given that many antibodies with different specificities have been characterized in the literature, future generalization of mBLM to additional antibody specificities will likely be achievable by extensive data mining (see discussion below). Besides, the continuous improvement of the speed of antibody discovery and characterization will also be beneficial, if not essential [3, 4].

The success of applying deep learning model to protein research can largely be attributed to the presence of databases such as Protein Data Bank (PDB) [59], UniProt [60], UniRef [61], 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 [62]. 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.

ACKNOWLEDGEMENTS

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.), and the Searle Scholars Program (N.C.W.). 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 help discussion.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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., I.R.S., D.C., Q.W.T. T.J.C.T., A.B.G. 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. The authors declare no other competing interests.

FIGURE LEGENDS

Figure 1. Germline gene usages in influenza HA antibodies. (A) The IGHV gene usage, **(B)** IGK(L)V gene usage, and **(C)** IGHD 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 Methods**).

Figure 2. Hydrophobicity of CDR H3 sequences. (A-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 1.5x inter-quartile range (IQR) of the third quartile, whereas the lower whisker extends to the lowest data point within 1.5x IQR of the first quartile. Each data point represents one antibody. The horizontal dotted line indicates the mean among antibodies from Genbank.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3. Antibody specificity prediction by memory B cell language model (mBLM). (A) Model architecture of 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)** Model performance of mBLM on the test set was evaluated by a normalized confusion matrix. **(C)** The performance of different antibody specificity prediction models was evaluated by F1 score, which represents the weighted harmonic mean of the precision and recall. CDR encoders: our previous model using a transformer encoder to encode CDR sequences [15]. ESM2: a general protein language model [18].

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, and H3 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 on to the structures of four antibodies in cluster 3 (PDB 4KVN [49], PDB 5KAQ [42], PDB 8GV6 [54], and PDB 3ZTJ [47]). The color scheme is same as that in panel A. **(E)** The relationship between saliency score and distance to the antigen (i.e. HA stem) is shown as a scatter plot. 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, 44-54].

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 5. Discovery of HA stem antibody by mBLM. (A-B) mBLM was applied to predict the specificity of **(A)** 60 antibodies to central stem epitope (left panel) and 38 to anchor stem epitope (right panel) that were reported recently [57], as well as **(B)** 4,452 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 18 HA unk antibodies that were predicted as HA stem antibodies was tested against mini-HA, which is an H1 stem-based construct [58]. Four known HA stem antibodies (051-09 5A02, 051-09 5E03, 310-18C3, and FI6v3) [47, 63, 64] were included as positive control. D2 H1-1/H3-1, which is a known HA head antibody [65], was included as negative control. In this binding experiment, antibodies were not purified from the supernatant and thus their concentrations were unknown. **(D)** Representative 2D classes from cryo-EM analysis of 310-18A5 Fab in complex with H1N1 A/Solomon Islands/3/2006 (SI06) HA are shown. Cyan arrows point to the 310-18A5 Fabs. **(E)** Cryo-EM 3D reconstruction of 310-18A5 Fab in complex with SI06 HA. Structural models of SI06 HA (PDB 6XSK) [66] and CR9114 (PDB 4FQH) [48] were docked into the 3D reconstruction.

METHODS

Collections of antibody information

Sequences of each human monoclonal antibody were from the original papers and/or NCBI GenBank database (**Table S1 and Table 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 [67, 68]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. [69, 70] 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 [69]. 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) [69], some antibody names from the second study did not (e.g. antibody name with prefix “20A-518-30”) [70]. As a result, 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 public clonotype

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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:

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

where WW represents the Wimley-White whole residue hydrophobicity scale [71] 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. Of note, upsampling only applied to the training set, but not the validation and test sets.

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) [72]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 test model generalization, the test set was generated with a maximum sequence identity of 80% to the training set. In other words, the pairwise sequence identity between individual antibody sequences in the test set and the training set was at least 20% (i.e. 26 amino acids). 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) [73] 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:

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We adapted RoBERTa [74] 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 β_1 = 0.9, β_2 = 0.999, ϵ = 10−8, 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

15-fold cross-validation

To evaluate the robustness of our mBLM for specificity prediction, we employed a 15-fold cross-validation approach for fine-tuning, specificity inference, and model explanation. We randomly down/upsampled and split the data 15 times, resulting in a diverse set of sequences in each iteration. Then, the model underwent 15 rounds of training and testing. For each iteration, model performance was evaluated. The overall model performance was quantified as the average across all iterations. The final predicted class represented the mode.

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Performance Metrics

The fine-tuned model was evaluated using the average F1 score, which represents the weighted harmonic mean of the precision and recall, as well as confusion matrix. The calculations were conducted using sklearn metrics functions [75].

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:

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

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 a $ReLU$ function:

$$S_i^c = ReLU\left(\alpha_i^c \mathrm{A}^i\right)$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where S_i^c represents the relative importance (saliency score) of residue i to specificity class c. The $ReLU$ 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 [76].

Sequence logo analysis

To identify sequence features within each cluster, we employed a thresholding approach based on the saliency scores. Specifically, for each cluster, we computed the frequency of each amino acid for residues with a saliency score >0.5. Then, sequence logos were generated by Logomaker in Python [77].

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).

Mammalian cell culture

HEK293T cells 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

552 using 0.05% Trypsin-EDTA solution (Gibco). Expi293F cells were maintained in Expi293
553 Expression Medium (Thermo Fisher Scientific). Sf9 cells (*Spodoptera frugiperda* ovarian cells,
554 female, ATCC) were maintained in Sf-900 II SFM medium (Thermo Fisher Scientific).
555
556 **Expression and purification of mini-HA and HA**
557 The mini-HA #4900 [58] and H1N1 A/Solomon Island/3/2006 HA were fused with N-terminal gp67
558 signal peptide and a C-terminal BirA biotinylation site, thrombin cleavage site, trimerization
559 domain, and a 6xHis-tag, and then cloned into a customized baculovirus transfer vector [46].
560 Subsequently, recombinant bacmid DNA was generated using the Bac-to-Bac system (Thermo
561 Fisher Scientific) according to the manufacturer's instructions. Baculovirus was generated by
562 transfecting the purified bacmid DNA into adherent Sf9 cells using Cellfectin reagent (Thermo
563 Fisher Scientific) according to the manufacturer's instructions. The baculovirus was further
564 amplified by passaging in adherent Sf9 cells at a multiplicity of infection (MOI) of 1. Recombinant
565 mini-HA protein was expressed by infecting 1 L of suspension Sf9 cells at an MOI of 1. On day 3
566 post-infection, Sf9 cells were pelleted by centrifugation at 4000 × *g* for 25 min, and soluble
567 recombinant mini-HA and HA were purified from the supernatant by affinity chromatography using
568 Ni Sepharose excel resin (Cytiva) and then size exclusion chromatography using a HiLoad 16/100
569 Superdex 200 prep grade column (Cytiva) in 20 mM Tris-HCl pH 8.0, 100 mM NaCl. The purified
570 mini-HA protein was concentrated by Amicon spin filter (Millipore Sigma) and filtered by 0.22 μm
571 centrifuge tube filters (Costar). Concentration of the protein was determined by nanodrop (Fisher
572 Scientific). Proteins were subsequent aliquoted, flash frozen by dry-ice ethanol mixture, and
573 stored at -80°C until used.
574
575 **Expression and purification of IgG**
576 The heavy and light chain genes of the obtained antibody were synthesized as eBlocks
577 (Integrated DNA Technologies), and then cloned into human IgG1 and human kappa or lambda

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

light chain expression vectors using Gibson assembly according to a previously described method [78]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

604 solution was added to each well. The absorbance of each well was measured at a wavelength of
605 450 nm using a Sunrise absorbance microplate reader (BioTek Synergy HTX Multimode Reader).
606
607 **Biolayer interferometry binding assay**
608 Binding assays were performed by biolayer interferometry (BLI) using an Octet Red96e
609 instrument (FortéBio) at room temperature as described previously [79]. Briefly, His-tagged mini-
610 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
611 Tween 20) were loaded onto anti-Penta-HIS (HIS1K) biosensors and incubated with the indicated
612 concentrations of Fabs. The assay consisted of five steps: (1) baseline: 60 s with 1× kinetics buffer;
613 (2) loading: 60 s with His-tagged mini-HA proteins; (3) baseline: 60 s with 1× kinetics buffer; (4)
614 association: 60 s with Fab samples; and (5) dissociation: 60 s with 1× kinetics buffer. For
615 estimating the exact K_D, a 1:1 binding model was used.
616
617 **Virus neutralization assay**
618 Madin-Darby canine kidney (MDCK) cells were seeded in a 96-well, flat-bottom cell culture plate
619 (Thermo Fisher). The next day, serially diluted monoclonal antibodies were mixed with an equal
620 volume of virus and incubated at 37°C for 1 hour. The antibody/virus mixture was then incubated
621 with the MDCK cells at 37°C after the cells were washed twice with PBS. Following a 1-hour
622 incubation, the antibody/virus mixture was replaced with Minimum Essential Medium (MEM)
623 supplemented with 25 mM of 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) and 1
624 µg mL^{-1} of Tosyl phenylalanyl chloromethyl ketone (TPCK)-trypsin. The plate was incubated at
625 37°C for 72 hours and the presence of virus was detected by hemagglutination assay. The results
626 were analyzed using Prism software (GraphPad).
627
628 **Cryogenic electron microscopy (cryo-EM) analysis**

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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).

Data availability

The cryoEM map of 310-18A5 Fab in complex with SI06 HA can be accessed at the Electron Microscopy Data Bank (EMDB) using accession code EMD-41849.

Model and code availability

Custom python scripts for all analyses and model training have been deposited to: https://github.com/nicwulab/HA_Abs.

REFERENCES

1. Graham BS, Gilman MSA, McLellan JS. Structure-based vaccine antigen design. Annu Rev Med. 2019;70:91-104. Epub 2019/01/30. doi: 10.1146/annurev-med-121217-094234. PubMed PMID: 30691364; PubMed Central PMCID: PMCPMC6936610.

2. Lu RM, Hwang YC, Liu IJ, Lee CC, Tsai HZ, Li HJ, et al. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020;27(1):1. Epub 2020/01/03. doi:

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

10.1186/s12929-019-0592-z. PubMed PMID: 31894001; PubMed Central PMCID: PMCPMC6939334.

3. Winters A, McFadden K, Bergen J, Landas J, Berry KA, Gonzalez A, et al. Rapid single B cell antibody discovery using nanopens and structured light. mAbs. 2019;11(6):1025-35. Epub 2019/06/13. doi: 10.1080/19420862.2019.1624126. PubMed PMID: 31185801; PubMed Central PMCID: PMCPMC6748590.

4. Curtis NC, Lee J. Beyond bulk single-chain sequencing: Getting at the whole receptor. Curr Opin Syst Biol. 2020;24:93-9. Epub 2020/10/27. doi: 10.1016/j.coisb.2020.10.008. PubMed PMID: 33102951; PubMed Central PMCID: PMCPMC7568503.

5. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566(7744):393-7. Epub 2019/01/22. doi: 10.1038/s41586-019-0879-y. PubMed PMID: 30664748; PubMed Central PMCID: PMCPMC6411386.

6. Schroeder HW, Jr. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Dev Comp Immunol. 2006;30(1-2):119-35. Epub 2005/08/09. doi: 10.1016/j.dci.2005.06.006. PubMed PMID: 16083957.

7. Wu NC, Yamayoshi S, Ito M, Uraki R, Kawaoka Y, Wilson IA. 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. 2018;24(4):569-78.e4. doi: 10.1016/j.chom.2018.09.010. PubMed PMID: 30308159.

8. Avnir Y, Tallarico AS, Zhu Q, Bennett AS, Connelly G, Sheehan J, et al. Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. PLoS Pathog. 2014;10(5):e1004103. doi: 10.1371/journal.ppat.1004103. PubMed PMID: 24788925; PubMed Central PMCID: PMCPMC4006906.

9. Zhou T, Lynch RM, Chen L, Acharya P, Wu X, Doria-Rose NA, et al. Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. Cell. 2015;161(6):1280-92. doi: 10.1016/j.cell.2015.05.007. PubMed PMID: 26004070; PubMed Central PMCID: PMCPMC4683157.

10. Robbiani DF, Bozzacco L, Keeffe JR, Khouri R, Olsen PC, Gazumyan A, et al. Recurrent potent human neutralizing antibodies to Zika virus in Brazil and Mexico. Cell. 2017;169(4):597-609.e11. Epub 2017/05/06. doi: 10.1016/j.cell.2017.04.024. PubMed PMID: 28475892; PubMed Central PMCID: PMCPMC5492969.

11. Ehrhardt SA, Zehner M, Krahling V, Cohen-Dvashi H, Kreer C, Elad N, et al. Polyclonal and convergent antibody response to Ebola virus vaccine rVSV-ZEBOV. Nat Med. 2019;25(10):1589-600. Epub 2019/10/09. doi: 10.1038/s41591-019-0602-4. PubMed PMID: 31591605.

12. Cohen-Dvashi H, Zehner M, Ehrhardt S, Katz M, Elad N, Klein F, et al. Structural basis for a convergent immune response against ebola virus. Cell Host Microbe. 2020;27(3):418-27.e4. Epub 2020/02/16. doi: 10.1016/j.chom.2020.01.007. PubMed PMID: 32059794.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

13. Chen EC, Gilchuk P, Zost SJ, Suryadevara N, Winkler ES, Cabel CR, et al. Convergent antibody responses to the SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Rep. 2021;36(8):109604. Epub 2021/08/20. doi: 10.1016/j.celrep.2021.109604. PubMed PMID: 34411541; PubMed Central PMCID: PMCPMC8352653.

14. Claireaux M, Caniels TG, de Gast M, Han J, Guerra D, Kerster G, et al. A public antibody class recognizes an S2 epitope exposed on open conformations of SARS-CoV-2 spike. Nat Commun. 2022;13(1):4539. Epub 2022/08/05. doi: 10.1038/s41467-022-32232-0. PubMed PMID: 35927266; PubMed Central PMCID: PMCPMC9352689 patent application related to this work filed by Amsterdam UMC (PCT/EP2021/062558, filed on 11 May 2021). The other authors declare that they have no competing interests.

15. Wang Y, Yuan M, Lv H, Peng J, Wilson IA, Wu NC. A large-scale systematic survey reveals recurring molecular features of public antibody responses to SARS-CoV-2. Immunity. 2022;55(6):1105-17.e4. Epub 2022/04/11. doi: 10.1016/j.immuni.2022.03.019. PubMed PMID: 35397794; PubMed Central PMCID: PMCPMC8947961.

16. Cia G, Pucci F, Rooman M. Critical review of conformational B-cell epitope prediction methods. Brief Bioinform. 2023;24(1):bbac567. Epub 2023/01/08. doi: 10.1093/bib/bbac567. PubMed PMID: 36611255.

17. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15):e2016239118. Epub 2021/04/21. doi: 10.1073/pnas.2016239118. PubMed PMID: 33876751; PubMed Central PMCID: PMCPMC8053943.

18. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123-30. Epub 2023/03/18. doi: 10.1126/science.ade2574. PubMed PMID: 36927031.

19. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102-10. Epub 2022/01/13. doi: 10.1093/bioinformatics/btac020. PubMed PMID: 35020807; PubMed Central PMCID: PMCPMC9386727.

20. Bordin N, Dallago C, Heinzinger M, Kim S, Littmann M, Rauer C, et al. Novel machine learning approaches revolutionize protein knowledge. Trends Biochem Sci. 2023;48(4):345-59. Epub 2022/12/13. doi: 10.1016/j.tibs.2022.11.001. PubMed PMID: 36504138.

21. Ferruz N, Schmidt S, Hocker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022;13(1):4348. Epub 2022/07/28. doi: 10.1038/s41467-022-32007-7. PubMed PMID: 35896542; PubMed Central PMCID: PMCPMC9329459.

22. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41:1099-106. Epub 2023/01/27. doi: 10.1038/s41587-022-01618-2. PubMed PMID: 36702895.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

23. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, et al. Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol. 2022;40(11):1617-23. Epub 2022/10/04. doi: 10.1038/s41587-022-01432-w. PubMed PMID: 36192636.

24. Shuai RW, Ruffolo JA, Gray JJ. Generative language modeling for antibody design. bioRxiv. 2022. doi: 10.1101/2021.12.13.472419.

25. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinform Adv. 2022;2(1):vbac046. Epub 2023/01/27. doi: 10.1093/bioadv/vbac046. PubMed PMID: 36699403; PubMed Central PMCID: PMCPMC9710568.

26. Hie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, Tang S, et al. Efficient evolution of human antibodies from general protein language models. Nat Biotechnol. 2023. Epub 2023/04/25. doi: 10.1038/s41587-023-01763-2. PubMed PMID: 37095349.

27. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics. 2021;37(5):734-5. Epub 2020/08/18. doi: 10.1093/bioinformatics/btaa739. PubMed PMID: 32805021; PubMed Central PMCID: PMCPMC7558925.

28. Wu NC, Wilson IA. A perspective on the structural and functional constraints for immune evasion: insights from influenza virus. J Mol Biol. 2017;429(17):2694-709. doi: 10.1016/j.jmb.2017.06.015. PubMed PMID: 28648617; PubMed Central PMCID: PMCPMC5573227.

29. Lang S, Xie J, Zhu X, Wu NC, Lerner RA, Wilson IA. 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. 2017;20(12):2935-43. doi: 10.1016/j.celrep.2017.08.084. PubMed PMID: 28930686.

30. Cheung CS, Fruehwirth A, Paparoditis PCG, Shen CH, Foglierini M, Joyce MG, et al. Identification and structure of a multidonor class of head-directed influenza-neutralizing antibodies reveal the mechanism for its recurrent elicitation. Cell Rep. 2020;32(9):108088. Epub 2020/09/03. doi: 10.1016/j.celrep.2020.108088. PubMed PMID: 32877670.

31. Wu NC, Yuan M, Liu H, Lee CD, Zhu X, Bangaru S, et al. An alternative binding mode of IGHV3-53 antibodies to the SARS-CoV-2 receptor binding domain. Cell Rep. 2020;33(3):108274. Epub 2020/10/08. doi: 10.1016/j.celrep.2020.108274. PubMed PMID: 33027617; PubMed Central PMCID: PMCPMC7522650.

32. Yuan M, Liu H, Wu NC, Lee CD, Zhu X, Zhao F, et al. Structural basis of a shared antibody response to SARS-CoV-2. Science. 2020;369(6507):1119-23. Epub 2020/07/15. doi: 10.1126/science.abd2321. PubMed PMID: 32661058; PubMed Central PMCID: PMCPMC7402627.

33. Wiley DC, Wilson IA, Skehel JJ. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature. 1981;289(5796):373-8. PubMed PMID: 6162101.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

34. Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell. 1982;31(2 Pt 1):417-27. PubMed PMID: 6186384.

35. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004;305(5682):371-6. doi: 10.1126/science.1097211. PubMed PMID: 15218094.

36. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science. 2013;342(6161):976-9. doi: 10.1126/science.1244730. PubMed PMID: 24264991.

37. Wu NC, Wilson IA. Influenza hemagglutinin structures and antibody recognition. Cold Spring Harb Perspect Med. 2020;10:a038778. Epub 2019/12/25. doi: 10.1101/cshperspect.a038778. PubMed PMID: 31871236.

38. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-D42. Epub 2012/11/30. doi: 10.1093/nar/gks1195. PubMed PMID: 23193287; PubMed Central PMCID: PMCPMC3531190.

39. Kovaltsuk A, Leem J, Kelm S, Snowden J, C.M. D, Krawczyk K. Observed Antibody Space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol. 2018;201:2502-9. PubMed PMID: 30217829.

40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv e-prints2016. p. arXiv:1610.02391.

41. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 2021;12(1):3168. Epub 2021/05/28. doi: 10.1038/s41467-021-23303-9. PubMed PMID: 34039967; PubMed Central PMCID: PMCPMC8155034.

42. Joyce MG, Wheatley AK, Thomas PV, Chuang GY, Soto C, Bailer RT, et al. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. Cell. 2016;166(3):609-23. doi: 10.1016/j.cell.2016.06.043. PubMed PMID: 27453470; PubMed Central PMCID: PMCPMC4978566.

43. Wu NC, Andrews SF, Raab JE, O'Connell S, Schramm CA, Ding X, et al. Convergent evolution in breadth of two V_H6-1-encoded influenza antibody clonotypes from a single donor. Cell Host Microbe. 2020;28:434-44. Epub 2020/07/04. doi: 10.1016/j.chom.2020.06.003. PubMed PMID: 32619441.

44. Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen LM, et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat Struct Mol Biol. 2009;16(3):265-73. doi: 10.1038/nsmb.1566. PubMed PMID: 19234466; PubMed Central PMCID: PMCPMC2692245.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, et al. Antibody recognition of a highly conserved influenza virus epitope. Science. 2009;324(5924):246-51. doi: 10.1126/science.1171491. PubMed PMID: 19251591; PubMed Central PMCID: PMCPMC2758658.

46. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, et al. A highly conserved neutralizing epitope on group 2 influenza A viruses. Science. 2011;333(6044):843-50. doi: 10.1126/science.1204839. PubMed PMID: 21737702; PubMed Central PMCID: PMCPMC3210727.

47. Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, Jarrossay D, et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science. 2011;333(6044):850-6. doi: 10.1126/science.1205669. PubMed PMID: 21798894.

48. Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R, Ekiert DC, et al. Highly conserved protective epitopes on influenza B viruses. Science. 2012;337(6100):1343-8. doi: 10.1126/science.1222908. PubMed PMID: 22878502; PubMed Central PMCID: PMCPMC3538841.

49. Nakamura G, Chai N, Park S, Chiang N, Lin Z, Chiu H, et al. An in vivo human-plasmablast enrichment technique allows rapid identification of therapeutic influenza A antibodies. Cell Host Microbe. 2013;14(1):93-103. doi: 10.1016/j.chom.2013.06.004. PubMed PMID: 23870317.

50. Friesen RH, Lee PS, Stoop EJ, Hoffman RM, Ekiert DC, Bhabha G, et al. A common solution to group 2 influenza virus neutralization. Proc Natl Acad Sci U S A. 2014;111(1):445-50. doi: 10.1073/pnas.1319058110. PubMed PMID: 24335589; PubMed Central PMCID: PMCPMC3890827.

51. Wu Y, Cho M, Shore D, Song M, Choi J, Jiang T, et al. A potent broad-spectrum protective human monoclonal antibody crosslinking two haemagglutinin monomers of influenza A virus. Nat Commun. 2015;6:7708. doi: 10.1038/ncomms8708. PubMed PMID: 26196962; PubMed Central PMCID: PMCPMC4518248.

52. Kallewaard NL, Corti D, Collins PJ, Neu U, McAuliffe JM, Benjamin E, et al. Structure and function analysis of an antibody recognizing all influenza A subtypes. Cell. 2016;166(3):596-608. doi: 10.1016/j.cell.2016.05.073. PubMed PMID: 27453466; PubMed Central PMCID: PMCPMC4967455.

53. Matsuda K, Huang J, Zhou T, Sheng Z, Kang BH, Ishida E, et al. Prolonged evolution of the memory B cell response induced by a replicating adenovirus-influenza H5 vaccine. Sci Immunol. 2019;4(34):eaau2710. Epub 2019/04/21. doi: 10.1126/sciimmunol.aau2710. PubMed PMID: 31004012.

54. Chen Y, Wang F, Yin L, Jiang H, Lu X, Bi Y, et al. Structural basis for a human broadly neutralizing influenza A hemagglutinin stem-specific antibody including H17/18 subtypes. Nat Commun. 2022;13(1):7603. Epub 2022/12/10. doi: 10.1038/s41467-022-35236-y. PubMed PMID: 36494358; PubMed Central PMCID: PMCPMC9734383.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

55. Benton DJ, Nans A, Calder LJ, Turner J, Neu U, Lin YP, et al. Influenza hemagglutinin membrane anchor. Proc Natl Acad Sci U S A. 2018;115(40):10112-7. Epub 2018/09/19. doi: 10.1073/pnas.1810927115. PubMed PMID: 30224494; PubMed Central PMCID: PMCPMC6176637.

56. Guthmiller JJ, Han J, Utset HA, Li L, Lan LY, Henry C, et al. Broadly neutralizing antibodies target a haemagglutinin anchor epitope. Nature. 2022;602(7896):314-20. Epub 2021/12/24. doi: 10.1038/s41586-021-04356-8. PubMed PMID: 34942633; PubMed Central PMCID: PMCPMC8828479.

57. Andrews SF, Cominsky LY, Shimberg GD, Gillespie RA, Gorman J, Raab JE, et al. An influenza H1 hemagglutinin stem-only immunogen elicits a broadly cross-reactive B cell response in humans. Sci Transl Med. 2023;15(692):eade4976. Epub 2023/04/19. doi: 10.1126/scitranslmed.ade4976. PubMed PMID: 37075126.

58. Impagliazzo A, Milder F, Kuipers H, Wagner MV, Zhu X, Hoffman RM, et al. A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. Science. 2015;349(6254):1301-6. doi: 10.1126/science.aac7263. PubMed PMID: 26303961.

59. ww PDBc. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47(D1):D520-D8. Epub 2018/10/26. doi: 10.1093/nar/gky949. PubMed PMID: 30357364; PubMed Central PMCID: PMCPMC6324056.

60. UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-D9. Epub 2020/11/26. doi: 10.1093/nar/gkaa1100. PubMed PMID: 33237286; PubMed Central PMCID: PMCPMC7778908.

61. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282-8. Epub 2007/03/24. doi: 10.1093/bioinformatics/btm098. PubMed PMID: 17379688.

62. Potocnakova L, Bhide M, Pulzova LB. An introduction to B-cell epitope mapping and in silico epitope prediction. J Immunol Res. 2016;2016:6760830. Epub 2017/01/28. doi: 10.1155/2016/6760830. PubMed PMID: 28127568; PubMed Central PMCID: PMCPMC5227168.

63. Andrews SF, Huang Y, Kaur K, Popova LI, Ho IY, Pauli NT, et al. Immune history profoundly affects broadly protective B cell responses to influenza. Sci Transl Med. 2015;7(316):316ra192. doi: 10.1126/scitranslmed.aad0522. PubMed PMID: 26631631; PubMed Central PMCID: PMCPMC4770855.

64. Whittle JR, Wheatley AK, Wu L, Lingwood D, Kanekiyo M, Ma SS, et al. Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. J Virol. 2014;88(8):4047-57. Epub 2014/02/07. doi: 10.1128/JVI.03422-13. PubMed PMID: 24501410; PubMed Central PMCID: PMCPMC3993745.

65. McCarthy KR, Lee J, Watanabe A, Kuraoka M, Robinson-McCarthy LR, Georgiou G, et al. A prevalent focused human antibody response to the influenza virus hemagglutinin head

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

895 interface. mBio. 2021;12(3):e0114421. Epub 2021/06/02. doi: 10.1128/mBio.01144-21.
896 PubMed PMID: 34060327; PubMed Central PMCID: PMCPMC8262862.

897 66. Moin SM, Boyington JC, Boyoglu-Barnum S, Gillespie RA, Cerutti G, Cheung CS, et al. Co-
898 immunization with hemagglutinin stem immunogens elicits cross-group neutralizing
899 antibodies and broad protection against influenza A viruses. Immunity. 2022;55(12):2405-
900 18.e7. Epub 2022/11/11. doi: 10.1016/j.immuni.2022.10.015. PubMed PMID: 36356572;
901 PubMed Central PMCID: PMCPMC9772109.

902 67. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain
903 sequence analysis tool. Nucleic Acids Res. 2013;41(Web Server issue):W34-W40. doi:
904 10.1093/nar/gkt382. PubMed PMID: 23671333; PubMed Central PMCID:
905 PMCPMC3692102.

906 68. Soto C, Finn JA, Willis JR, Day SB, Sinkovits RS, Jones T, et al. PyIR: a scalable wrapper
907 for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST.
908 BMC Bioinformatics. 2020;21(1):314. Epub 2020/07/18. doi: 10.1186/s12859-020-03649-5.
909 PubMed PMID: 32677886; PubMed Central PMCID: PMCPMC7364545.

910 69. Andrews SF, Joyce MG, Chambers MJ, Gillespie RA, Kanekiyo M, Leung K, et al.
911 Preferential induction of cross-group influenza A hemagglutinin stem-specific memory B
912 cells after H7N9 immunization in humans. Sci Immunol. 2017;2(13):eaan2676. doi:
913 10.1126/sciimmunol.aan2676. PubMed PMID: 28783708.

914 70. Andrews SF, Chambers MJ, Schramm CA, Plyler J, Raab JE, Kanekiyo M, et al. Activation
915 dynamics and immunoglobulin evolution of pre-existing and newly generated human
916 memory B cell responses to influenza hemagglutinin. Immunity. 2019;51(2):398-410.e5.
917 Epub 2019/07/28. doi: 10.1016/j.immuni.2019.06.024. PubMed PMID: 31350180.

918 71. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at
919 membrane interfaces. Nat Struct Biol. 1996;3(10):842-8. Epub 1996/10/01. doi:
920 10.1038/nsb1096-842. PubMed PMID: 8836100.

921 72. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or
922 nucleotide sequences. Bioinformatics. 2006;22(13):1658-9. Epub 2006/05/30. doi:
923 10.1093/bioinformatics/btl158. PubMed PMID: 16731699.

924 73. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional
925 Transformers for Language Understanding. arXiv. 2018. doi: 10.48550/arXiv.1810.04805.

926 74. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT
927 Pretraining Approach. arXiv. 2019. doi: 10.48550/arXiv.1907.11692.

928 75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
929 machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825-30.

930 76. Waskom M. seaborn: statistical data visualization. Journal of Open Source Software.
931 2021;6(60):3021. doi: 10.21105/joss.03021.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.11.557288; this version posted September 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

932 77. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics.
933 2020;36(7):2272-4. Epub 2019/12/11. doi: 10.1093/bioinformatics/btz921. PubMed PMID:
934 31821414; PubMed Central PMCID: PMCPMC7141850.

935 78. Guthmiller JJ, Dugan HL, Neu KE, Lan LY, Wilson PC. An efficient method to generate
936 monoclonal antibodies from human B cells. Methods Mol Biol. 2019;1904:109-45. Epub
937 2018/12/13. doi: 10.1007/978-1-4939-8958-4_5. PubMed PMID: 30539468.

938 79. Wu NC, Grande G, Turner HL, Ward AB, Xie J, Lerner RA, et al. In vitro evolution of an
939 influenza broadly neutralizing antibody is modulated by hemagglutinin receptor specificity.
940 Nature Communications. 2017;8(1):15371.
941

Figure 1

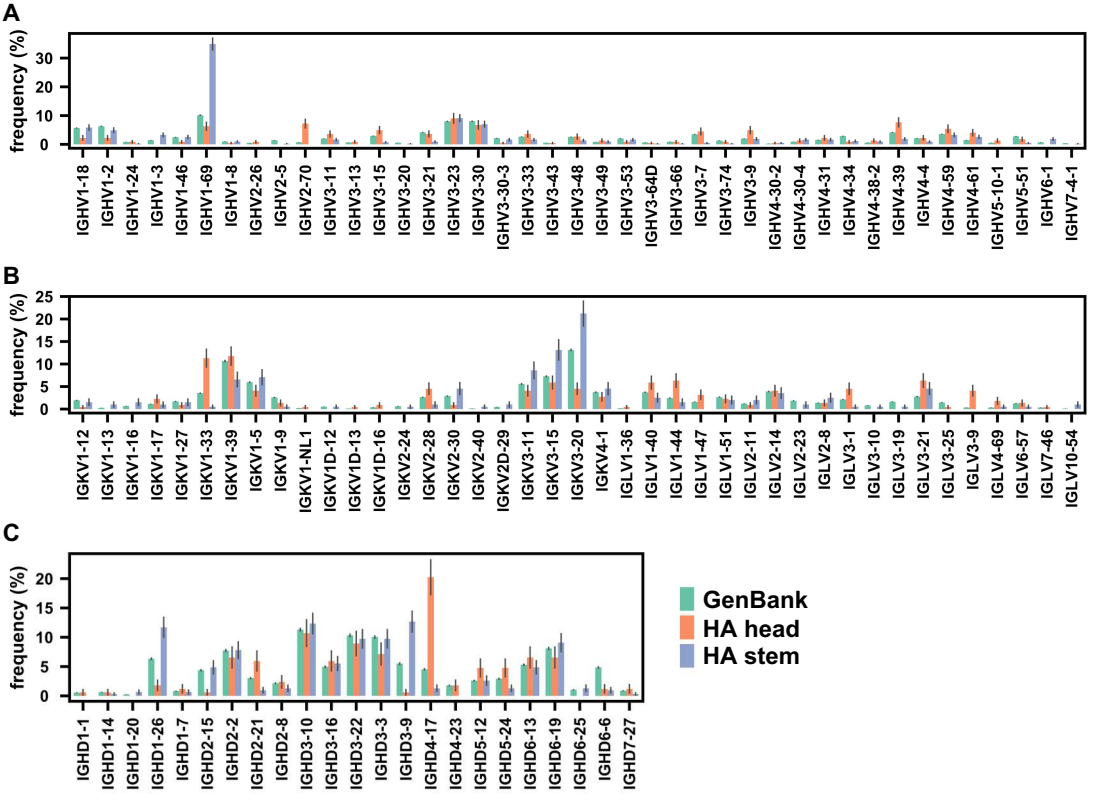

A
frequency (%)
0
10
20
30
IGHV1-18
IGHV1-2
IGHV1-24
IGHV1-3
IGHV1-46
IGHV1-69
IGHV1-8
IGHV2-26
IGHV2-5
IGHV2-70
IGHV3-11
IGHV3-13
IGHV3-15
IGHV3-20
IGHV3-21
IGHV3-23
IGHV3-30
IGHV3-30-3
IGHV3-33
IGHV3-43
IGHV3-48
IGHV3-49
IGHV3-53
IGHV3-64D
IGHV3-66
IGHV3-7
IGHV3-74
IGHV3-9
IGHV4-30-2
IGHV4-30-4
IGHV4-31
IGHV4-34
IGHV4-38-2
IGHV4-39
IGHV4-4
IGHV4-59
IGHV4-61
IGHV5-10-1
IGHV5-51
IGHV6-1
IGHV7-4-1
B
frequency (%)
0
5
10
15
20
25
IGKV1-12
IGKV1-13
IGKV1-16
IGKV1-17
IGKV1-27
IGKV1-33
IGKV1-39
IGKV1-5
IGKV1-9
IGKV1-NL1
IGKV1D-12
IGKV1D-13
IGKV1D-16
IGKV2-24
IGKV2-28
IGKV2-30
IGKV2-40
IGKV2D-29
IGKV3-11
IGKV3-15
IGKV3-20
IGKV4-1
IGLV1-36
IGLV1-40
IGLV1-44
IGLV1-47
IGLV1-51
IGLV2-11
IGLV2-14
IGLV2-23
IGLV2-8
IGLV3-1
IGLV3-10
IGLV3-19
IGLV3-21
IGLV3-25
IGLV3-9
IGLV4-69
IGLV6-57
IGLV7-46
IGLV10-54
C
frequency (%)
0
5
10
15
20
IGHD1-1
IGHD1-14
IGHD1-20
IGHD1-26
IGHD1-7
IGHD2-15
IGHD2-2
IGHD2-21
IGHD2-8
IGHD3-10
IGHD3-16
IGHD3-22
IGHD3-3
IGHD3-9
IGHD4-17
IGHD4-23
IGHD5-12
IGHD5-24
IGHD6-13
IGHD6-19
IGHD6-25
IGHD6-6
IGHD7-27
GenBank
HA head
HA stem

Figure 2

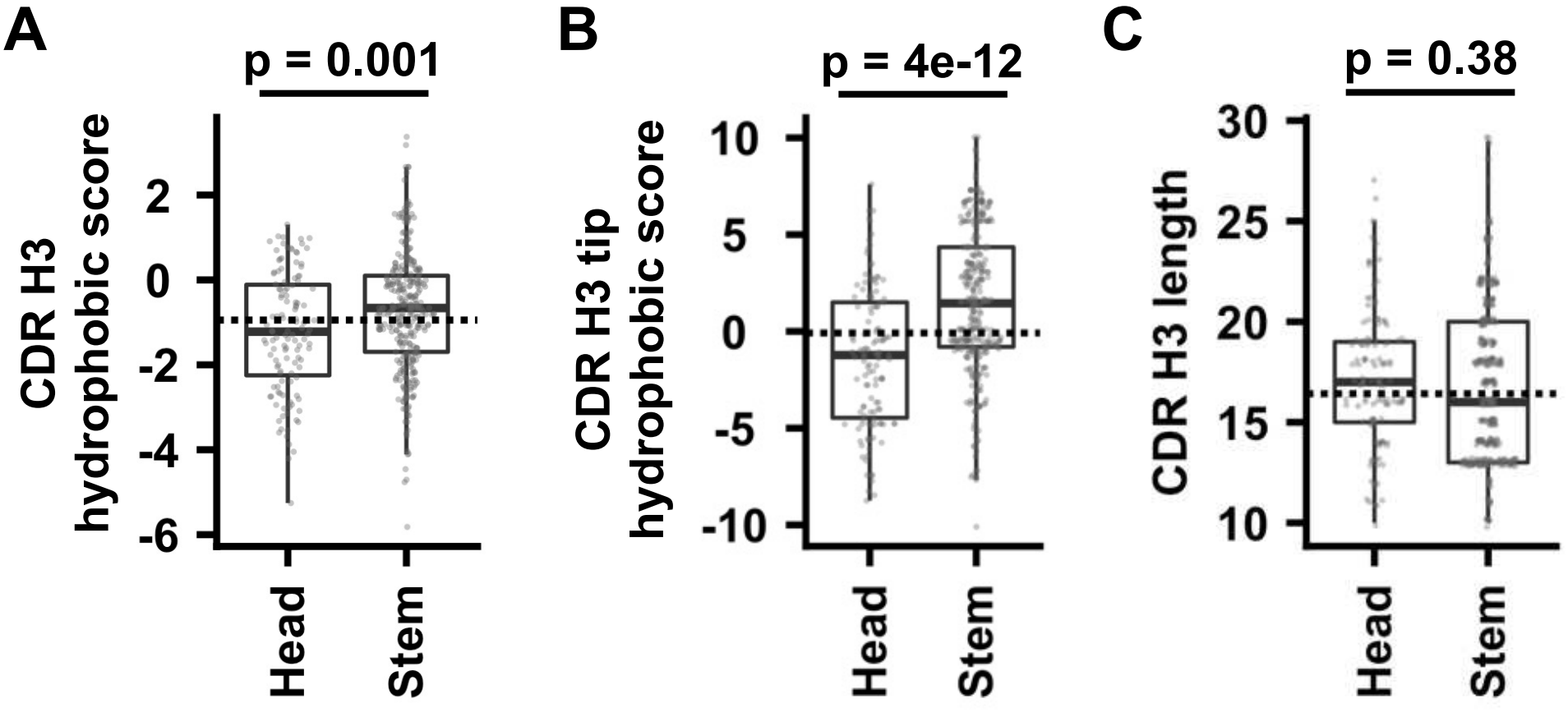
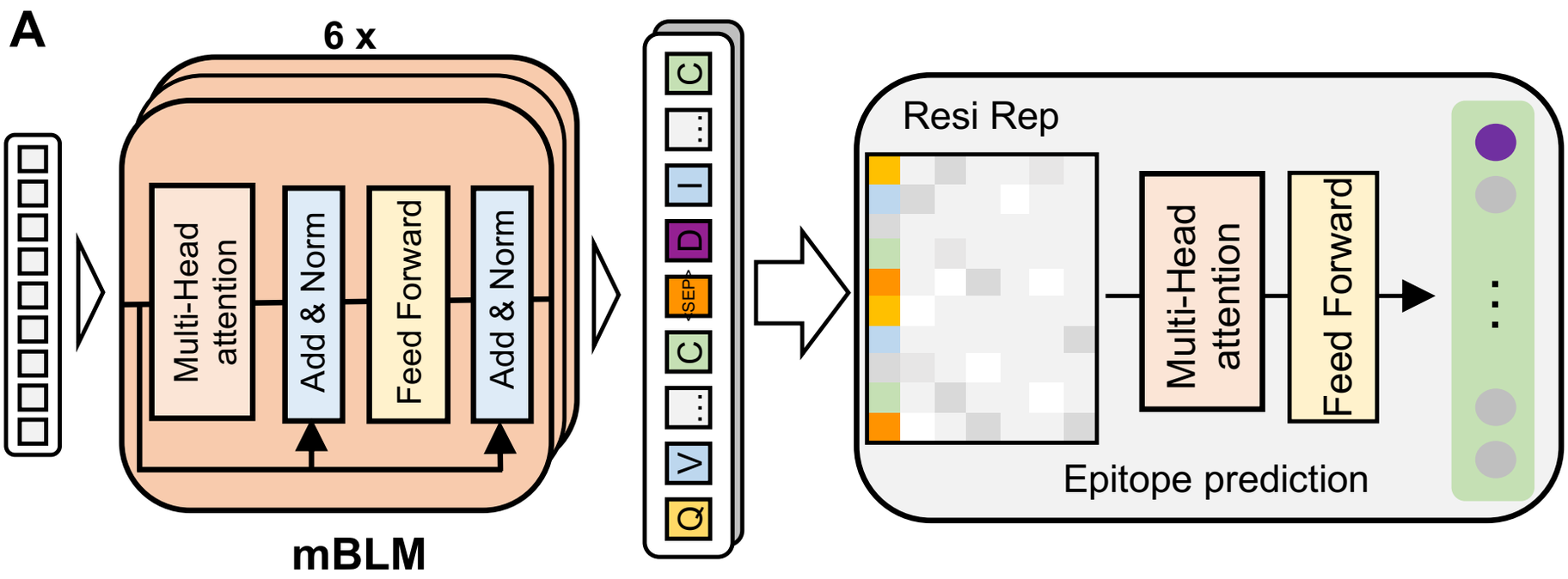
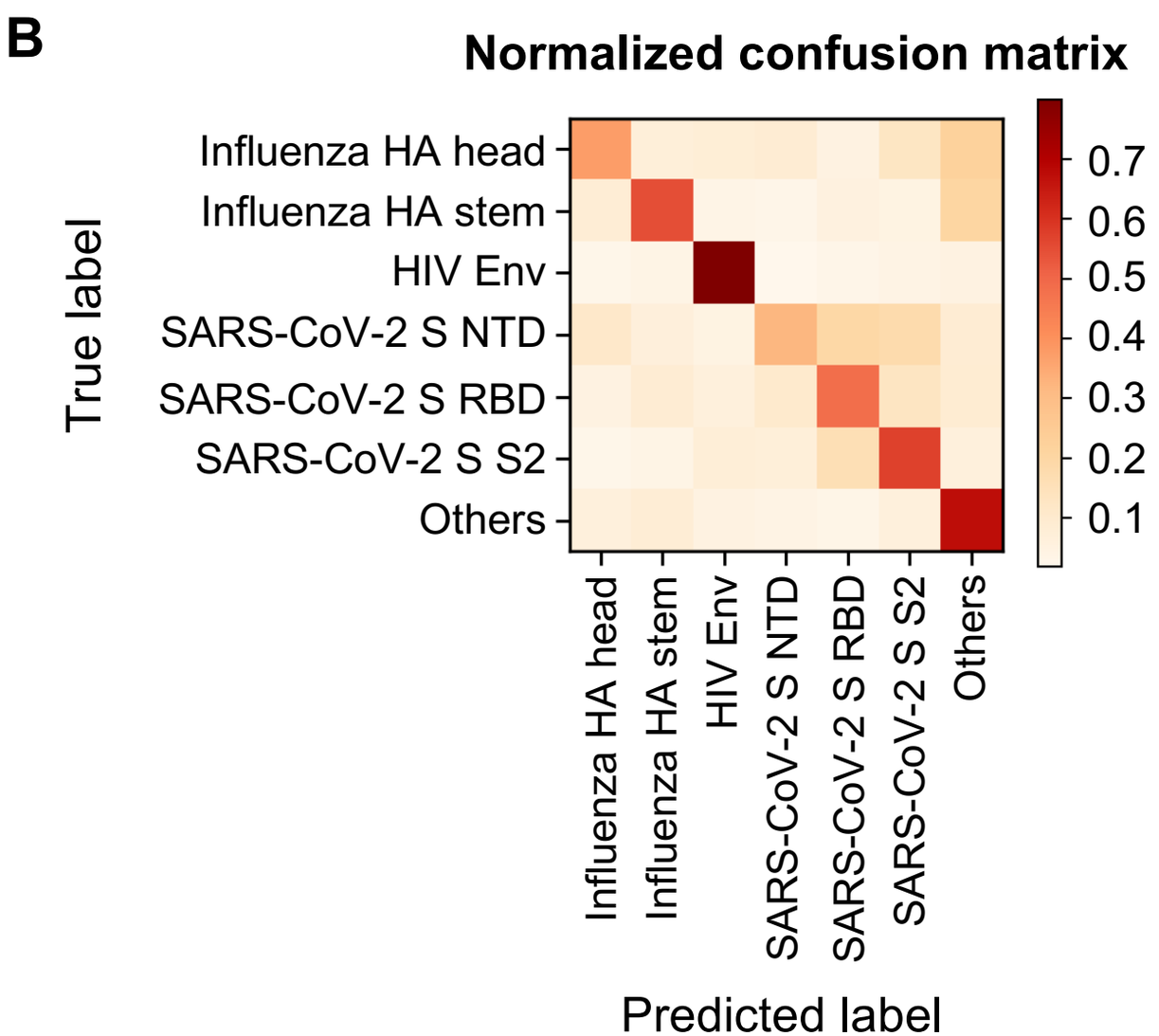

A
p = 0.001
CDR H3 hydrophobic score
2
0
-2
-4
-6
Head
Stem
B
p = 4e-12
CDR H3 tip hydrophobic score
10
5
0
-5
-10
Head
Stem
C
p = 0.38
CDR H3 length
30
25
20
15
10
Head
Stem

Figure 3


A
6 x
Multi-Head attention
Add & Norm
Feed Forward
Add & Norm
mBLM
C
...
I
D
SEP
C
...
V
Q
Resi Rep
Multi-Head attention
Feed Forward
Epitope prediction


B
Normalized confusion matrix
True label
Influenza HA head
Influenza HA stem
HIV Env
SARS-CoV-2 S NTD
SARS-CoV-2 S RBD
SARS-CoV-2 S S2
Others
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Predicted label

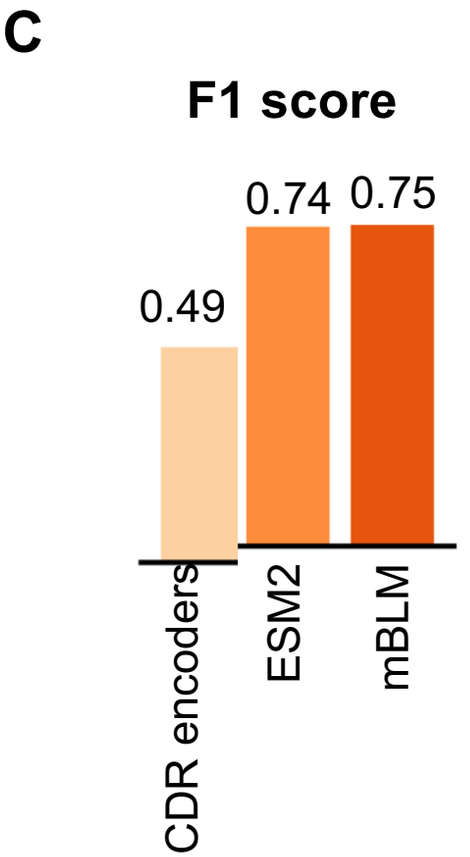

C
F1 score
0.49
0.74
0.75
CDR encoders
ESM2
mBLM

Figure 4

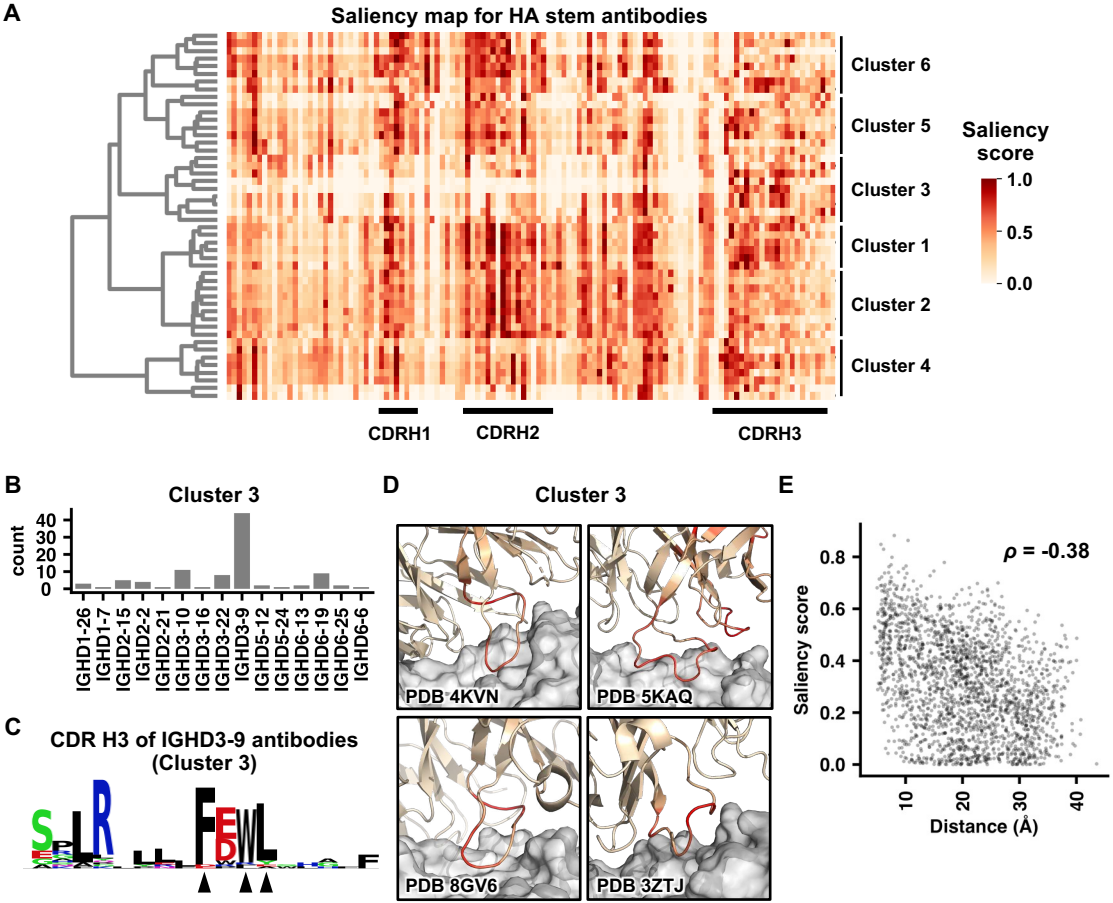
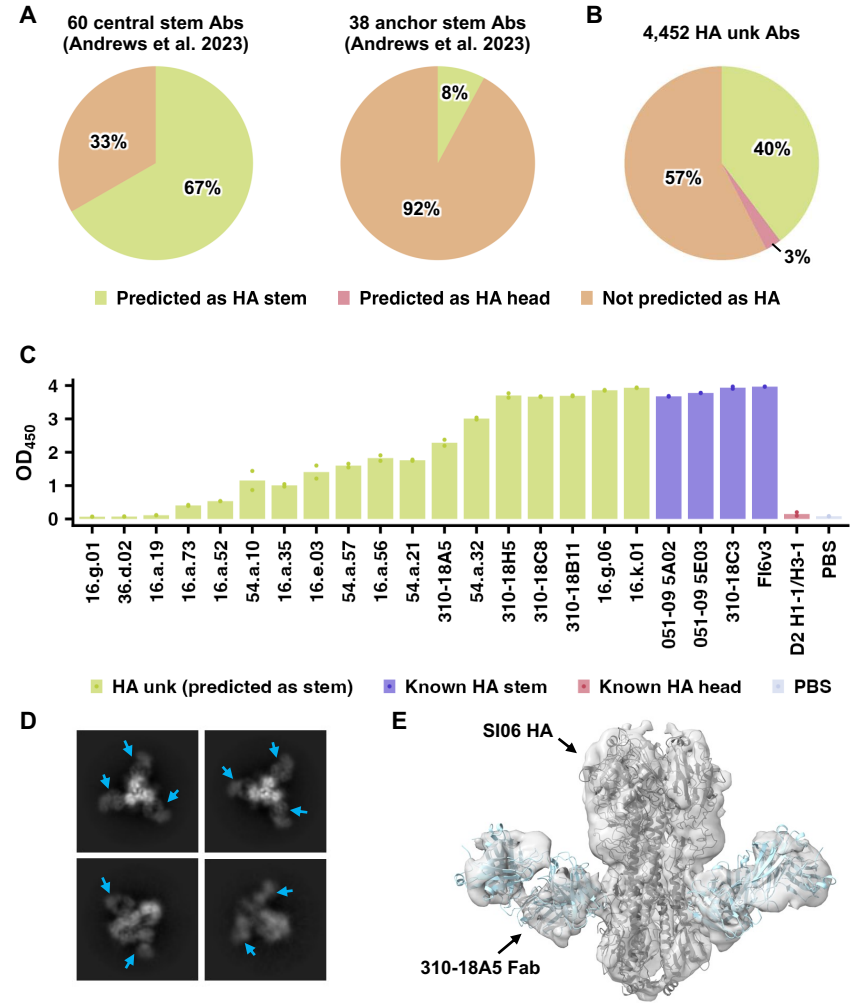

A
Saliency map for HA stem antibodies
Cluster 6
Cluster 5
Cluster 3
Cluster 1
Cluster 2
Cluster 4
Saliency score
1.0
0.5
0.0
CDRH1
CDRH2
CDRH3
B
Cluster 3
count
40
30
20
10
0
IGHD1-26
IGHD1-7
IGHD2-15
IGHD2-2
IGHD2-21
IGHD3-10
IGHD3-16
IGHD3-22
IGHD3-9
IGHD5-12
IGHD5-24
IGHD6-13
IGHD6-19
IGHD6-25
IGHD6-6
C
CDR H3 of IGHD3-9 antibodies
(Cluster 3)
D
Cluster 3
PDB 4KVN
PDB 5KAQ
PDB 8GV6
PDB 3ZTJ
E
ρ = -0.38
Saliency score
0.8
0.6
0.4
0.2
0.0
10
20
30
40
Distance (Å)

Figure 5


A
60 central stem Abs
(Andrews et al. 2023)
33%
67%
38 anchor stem Abs
(Andrews et al. 2023)
8%
92%
B
4,452 HA unk Abs
40%
57%
3%
Predicted as HA stem
Predicted as HA head
Not predicted as HA
C
OD450
0
1
2
3
4
16.g.01
36.d.02
16.a.19
16.a.73
16.a.52
54.a.10
16.a.35
16.e.03
54.a.57
16.a.56
54.a.21
310-18A5
54.a.32
310-18H5
310-18C8
310-18B11
16.g.06
16.k.01
051-09 5A02
051-09 5E03
310-18C3
FI6v3
D2 H1-1/H3-1
PBS
HA unk (predicted as stem)
Known HA stem
Known HA head
PBS
D
E
SI06 HA
310-18A5 Fab