











RAIN: machine learning-based identification for HIV-1 bNAbs

Received: 29 February 2024

Accepted: 17 June 2024

Published online: 24 June 2024

 Check for updates

Mathilde Foglierini ^{1,2,3,11}, Pauline Nortier^{1,2,11}, Rachel Schelling^{1,2}, Rahel R. Winiger ^{1,2}, Philippe Jacquet ⁴, Sijy O'Dell⁵, Davide Demurtas⁶, Maxmillian Mpina⁷, Omar Lweno ⁷, Yannick D. Muller ^{1,2}, Constantinos Petrovas⁸, Claudia Daubenberger ^{9,10}, Matthieu Perreau¹, Nicole A. Doria-Rose ⁵, Raphael Gottardo ³ & Laurent Perez ^{1,2} 

Broadly neutralizing antibodies (bNAbs) are promising candidates for the treatment and prevention of HIV-1 infections. Despite their critical importance, automatic detection of HIV-1 bNAbs from immune repertoires is still lacking. Here, we develop a straightforward computational method for the Rapid Automatic Identification of bNAbs (RAIN) based on machine learning methods. In contrast to other approaches, which use one-hot encoding amino acid sequences or structural alignment for prediction, RAIN uses a combination of selected sequence-based features for the accurate prediction of HIV-1 bNAbs. We demonstrate the performance of our approach on non-biased, experimentally obtained and sequenced BCR repertoires from HIV-1 immune donors. RAIN processing leads to the successful identification of distinct HIV-1 bNAbs targeting the CD4-binding site of the envelope glycoprotein. In addition, we validate the identified bNAbs using an *in vitro* neutralization assay and we solve the structure of one of them in complex with the soluble native-like heterotrimeric envelope glycoprotein by single-particle cryo-electron microscopy (cryo-EM). Overall, we propose a method to facilitate and accelerate HIV-1 bNAbs discovery from non-selected immune repertoires.

More than 40 years after its identification, the human immunodeficiency virus-1 (HIV-1) remains a major global health concern¹. The World Health Organization (WHO) estimates 38 million HIV-1 infected individuals worldwide in 2023, 1.5 million new HIV-1 infections, and 650,000 deaths from acquired immunodeficiency syndrome (AIDS)-related illness. Despite intense research efforts, there is still no cure nor vaccine for HIV-1 infections available². Humoral immune response

to HIV-1 targets the envelope (Env) protein of the virion, a trimeric membrane glycoprotein complex comprising gp120 and gp41³. However, the virus rapidly escapes immune control due to the exceptional Env glycoprotein diversity generated by the error-prone replication machinery of HIV-1⁴. Moreover, additional mechanisms of immune evasion exist, such as heavy glycosylation of gp120, which promotes a conformational masking of the receptor-binding site⁵. Screening of

¹Department of Medicine, Service of Immunology and Allergy, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland. ²Centre for Human Immunology, Lausanne, Switzerland. ³Biomedical Data Science Centre, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland. ⁴Scientific Computing and Research Support Unit, University of Lausanne, Lausanne, Switzerland. ⁵Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ⁶Interdisciplinary center of electron microscopy, CIME, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ⁷Ifakara Health Institute, Bagamoyo, United Republic of Tanzania. ⁸Department of Laboratory Medicine and Pathology, Institute of Pathology, Lausanne University Hospital, Lausanne, Switzerland. ⁹Department of Medical Parasitology and Infection Biology, Clinical Immunology Unit, Swiss Tropical and Public Health Institute, Basel, Switzerland. ¹⁰University of Basel, Basel, Switzerland. ¹¹These authors contributed equally: Mathilde Foglierini, Pauline Nortier. ✉ e-mail: laurent.perez@chuv.ch

plasma from HIV-1 seropositive (HIV-1+) subjects led to the identification of rare individuals possessing sera with broad and potent neutralizing activities against numerous HIV-1 viruses. Additional studies allowed the cloning and sequencing of B-cell receptors (BCRs) and permitted the identification of broadly neutralizing antibodies (bNAb)s, which can neutralize most viral strains at low concentrations *in vitro*⁶. Investigation of the development and structural properties of these bNAb)s revealed only a low level of sequence identity between them, but demonstrated that specific characteristics are associated with their function. For example, bNAb)s exhibit an extreme level of somatic hypermutations (SHMs) and large nucleotide insertions leading to long heavy chain complementary determining regions (CDRs)^{7,8}.

Since their identification, bNAb)s have gained intense therapeutic interest. Although approved drugs against HIV-1 infection exist, passive antibody prophylaxis and immunotherapy could hold a valuable place in both prevention and treatment⁹. Passive transfer of bNAb)s demonstrated a decrease of viral loads^{10,11}, prevention of infection^{12,13}, delay of viral rebound^{14,15}, and suppression of viremia in humanized mice, non-human primates, and humans without notable adverse events or side effects^{16,17}. bNAb)s target distinct sites of vulnerability at the surface of the envelope: the CD4-binding site (CD4bs), variable loop V1/V2 apex, and V3 loop, a larger site spanning the interface between gp41 and gp120 (interface) including the fusion peptide, and the membrane-proximal external region (MPER). Recently, a sixth site was discovered, defined by the bNAb VRC-PG05, which binds to the center of the so-called “silent face” of gp120¹⁸.

To date, the identification of bNAb)s has required B-cell isolation and clonal expansion from selected individuals possessing sera with broadly neutralizing activity. This step is followed by antibody cloning and experimental validation of their neutralization potential. While both steps represent an important research effort, the process has benefited from identified immune donors¹⁹ and the development of high-throughput analyses of antibody repertoires by next-generation sequencing (NGS). Still, the number of identified HIV bNAb)s remains relatively low, with only 255 of them being reported^{3,20}. Some bNAb)s have been investigated in registered clinical trials, for prevention, as a component of long-acting antiretroviral therapy (ART), or intervention aimed at long-term drug-free remission of HIV^{17,21,22}. However, the clinical success of bNAb) passive immunization strategies will likely require a combination of antibodies to increase the overall breadth and potency against diverse HIV-1 isolates and to prevent the emergence of resistance²³. The recent deployment of large datasets of human B-cell repertoires on database repositories represents an opportunity for novel bNAb) identification assuming that computational tools for their automatic identification and classification are developed²⁴. Artificial intelligence (AI)-based prediction tools to find the antibodies and antigens have been developed²⁵. However, most of these tools rely on structural or amino acid sequence similarities of related antibodies to identify potential target proteins²⁶. Nonetheless, despite important research and characterization efforts, a precise set of criteria required for classifying bNAb)s versus non-bNAb)s is still lacking.

Here, we developed a computational pipeline named RAIN for the Rapid Automatic Identification of bNAb)s from Immune Repertoires. RAIN is based on four different machine-learning algorithms, which can be trained in just a few minutes using a Python script. RAIN only requires the following: a cellranger scBCR output going through the Immcantation pipeline, and a R script converting the repertoire data into a features table for bNAb) prediction. We validated RAIN on previously identified bNAb)s, leading to a prediction accuracy of 100% and an Area Under the Curve (AUC) value ranging from 0.92 to 1, depending on the antigenic site. In addition, we isolated class-switched memory B cells from HIV-1 immune donors and performed single-cell BCR sequencing to demonstrate the method's performance. Importantly, immune repertoire analysis of donors with a serum able to broadly neutralize different HIV-1 isolates led to the identification of three bNAb)s, while none was detected

in the repertoire of immune donors with sera that did not possess broadly neutralizing activities. The identified bNAb)s were further assessed for their affinities to the stabilized envelope prefusion trimer BG505 DS-SOSIP, for their neutralizing activities and one of them was additionally characterized by cryo-EM.

Results

Subset of discrete characteristics discriminates HIV-1 bNAb)s from mAb)s

The automatic identification of HIV bNAb)s cannot be solely based on amino acid sequence similarity of the heavy or light chains, due to a large sequence variability resulting from the long affinity maturation process. In contrast, HIV-1 bNAb)s isolated from chronically infected adults exhibit a signature of characteristic features, including high SHMs, insertions or deletions (indels), long complementarity-determining regions H3 (CDRH3), high potency, and broad viral neutralization breadth³. Moreover, the VRC01-class bNAb)s, targeting the CD4bs, have also been shown to preferentially use specific germline alleles^{27,28} and possess an unusually short CDRL3 of only five amino acids. These short CDRL3 are essential to contact gp120, while avoiding the glycan at position N267 in the D loop of gp120²⁹. While bNAb)s targeting the V1V2 apex use specific IGHV genes and together with bNAb)s binding the V3 glycan, they are characterized by a long (20–34 residues) CDRH3 sequence^{30,31}. We hypothesized that integrating specific parameters characterizing HIV-1 bNAb)s in a machine-learning framework could allow a rapid identification of bNAb)s from an immune repertoire (Fig. 1). To identify predictors of HIV-1 bNAb)s, we investigated specific features associated with these antibodies and inferred them from their highly diversified amino acid sequences. We collected and curated bNAb) sequences from the CATNAP (Compile, Analyze, and Tally NAb) Panels) database³². Data curation consisted of only considering human affinity matured sequences and removing incomplete or unpaired sequences (Supplementary Data file 1). We obtained a total of 255 bNAb) paired sequences, described to bind the V1V2 apex ($n = 98$), V3 glycan ($n = 56$), CD4-binding site ($n = 54$), gp120/gp41 interface ($n = 26$), and MPER ($n = 21$). To visualize the sequence similarity among these selected bNAb)s, we initially aligned the sequences using ANARCI with the IMGT format³³. Subsequently, we computed an identity score matrix to represent the inverse of the Levenshtein distance for each sequence pair. We selected either the full-length variable heavy (VH) or only the CDRH3 amino acid sequence (Fig. 2a). As expected, VH and CDRH3 share only minimal conservation among HIV-1 bNAb)s. This result indicates that a homology and alignment approach to identify bNAb)s would probably be unsuccessful. Next, to create a dataset of paired BCR sequences that is unlikely to recognize an HIV antigen (hereafter named mAb)s, we retrieved and curated paired antibody sequences from ten healthy seronegative donors to obtain a total of 14,962 sequences (Supplementary Table 1). To control the comparability of the HIV-1 bNAb)s with unassigned mAb)s, we performed an additional similarity matrix with VH and CDRH3 comparing bNAb)s and curated mAb)s (Fig. 2b). As anticipated, only low similarity levels were observed, a result in agreement with the precedent matrix and indicating that the sequences were amenable to machine-learning approaches.

We then decided to investigate if some of the bNAb)s distinct properties could be used as predictive variables for each targeted antigenic site. We considered as potential predictors the length of the CDR3 for the heavy (H3) and light (L3) chains, the frequency of SHM in the V gene (v) or unconventional acquired mutations in the framework regions only (uv), and the hydrophobicity of CDRH3^{34,35} (φ) (Fig. 3a–e). Interestingly, anti-CD4bs bNAb)s analysis demonstrated a statistically higher SHM frequency, a higher frequency of unconventional mutations (outside of the CDRs)³⁵, and a significantly shorter length of CDRL3 (Fig. 3a, b, e and Supplementary Fig. 1a) compared to the control mAb)s reported in Supplementary Table 1. For the anti-MPER

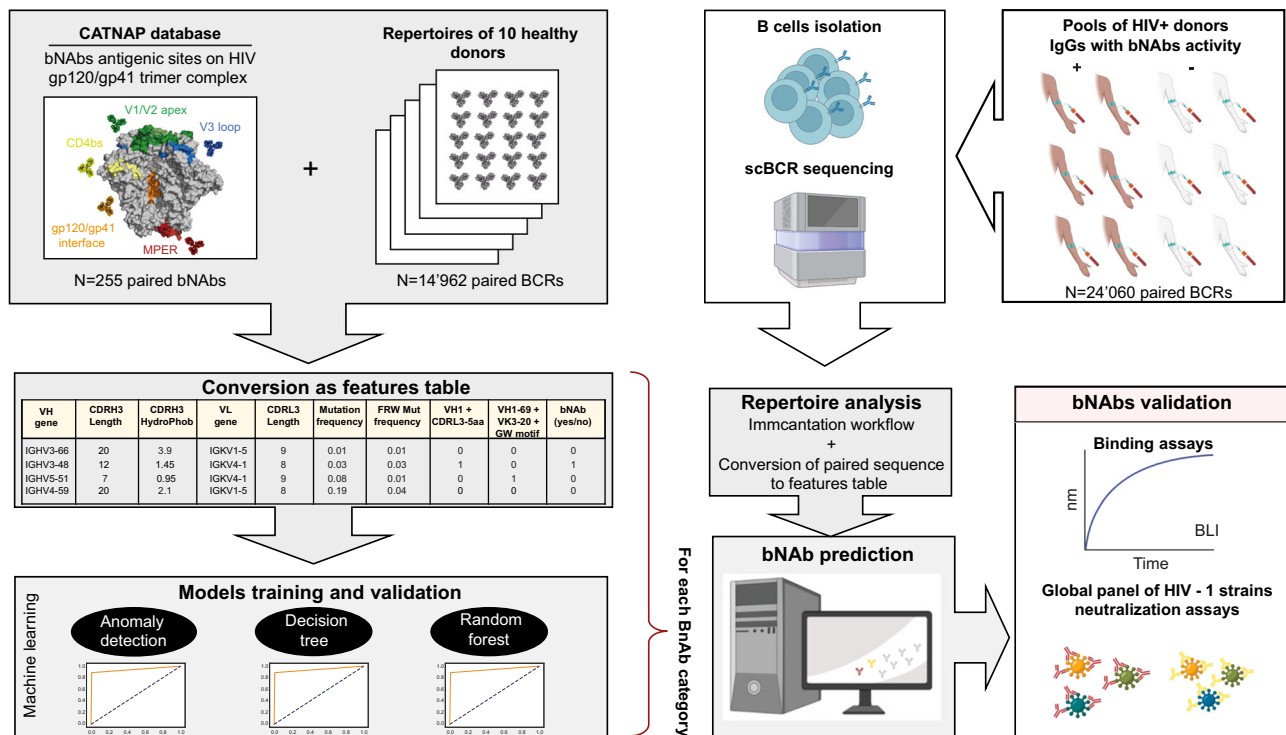


Fig. 1 | RAIN pipeline for automatic identification of bNAbs. Data collected from the CATNAP database (bNAbs) and healthy donor repertoires (mAbs) were converted into a features table to train and validate four machine-learning models: anomaly detection (AD), Decision Tree (DT), random forest (RF), and super learner (SL). We performed single-cell BCR sequencing from HIV-1 seropositive donors with

sera exhibiting broadly neutralizing activities (illustrated by the brown arms) or without neutralizing activities (illustrated by the white arms). BCR sequences were processed by the Immcatation workflow and analyzed as a feature table. Next, the predicted bNAbs found by the four algorithms were produced and tested in neutralization and binding assays. The image was created using BioRender.com.

bNAbs, we observed a longer CDRH3, with higher hydrophobicity, and a higher mutation frequency in both V gene and framework (FRW) regions (Fig. 3a–d and Supplementary Fig. 1b). The bNAbs targeting the V1V2 apex showed a higher mutation frequency of the V gene, but the difference was mainly due to a higher hydrophobicity of the CDRH3 and a longer CDRH3 (Fig. 3a–d and Supplementary Fig. 1c). bNAbs targeting the V3 glycan have a higher mutation frequency, slightly higher hydrophobicity of the CDRH3 and a longer CDRH3 (Fig. 3a–d and Supplementary Fig. 1d). bNAbs targeting the interface region demonstrated an increased frequency of mutations in the V gene and FRW regions (Fig. 3a, b and Supplementary Fig. 1e). Part of these results were expected but confirmed that this set of characteristics is statistically different between bNAbs and mAbs. To further investigate if these characteristics could be used to discriminate between bNAbs and mAbs, we decided to use them as variables in a two-dimensional Principal Component Analysis (PCA) (Fig. 3f–j). Remarkably, the five characteristics were sufficient to separate bNAbs from mAbs into two distinct clusters within each category of antigenic sites. We observed an explained variation of 0.43 for PC1 and 0.29 for PC2 across all five antigenic sites, while the weights of the features exhibited striking similarities. For PC1, the frequency of mutation in both, the CDRs and framework regions were important, whereas the hydrophobicity and length of CDRH3 were important for PC2. Unexpectedly, the length of CDRL3 was a less important feature. Based on these observations, we decided to use this set of measurable characteristics as predictors to distinguish bNAbs from mAbs.

Algorithm selection and validation for the computational pipeline

To further investigate the feasibility of automatic identification of potential HIV-1 bNAbs, we decided to use different machine-learning approaches to increase robustness and decrease the likelihood of false

predictions. First, antibody sequences were converted into a list of values corresponding to the set of predictors identified previously. bNAb sequences coming from the CATNAP database were annotated using Igbblast and the Immcatation workflow^{36–38}. The resulting Adaptive Immune Receptor Repertoire (AIRR) characteristics were converted to a feature format table. Similarly, mAb sequences obtained from public databases were processed as described previously³⁹ and converted into a features table. For each antigenic site, bNAbs and mAbs were pooled as one dataset and subdivided into three: 60% as a training set and 20% each as a validation and test set, respectively. An anomaly detection (AD) algorithm has been used in the specific case of a binary classification task, where one group appears as an outlier⁴⁰. Given the scarcity of reported HIV-1 bNAbs compared to the quantity of mAbs, we first opted for the AD algorithm to automatically identify bNAbs. We used the multivariate Gaussian model based on a threshold value (Epsilon) to estimate the probability of an antibody being flagged as ‘anomaly’ or not. Then, the optimal Epsilon parameter minimizing the number of false positives was obtained using the validation set (Supplementary Fig. 2a–e and Supplementary Table 2), while the evaluation of the AD performance, including computing of the area under the curve (AUC) was done with the test set (Fig. 4a, b). We observed that the AD algorithm discriminates well bNAbs targeting the V1V2 apex (AUC: 0.93), the CD4bs (AUC: 0.88), the MPER (AUC: 0.82), and the interface (AUC: 0.8). However, bNAbs targeting the V3 glycan were poorly identified, with an AUC of 0.64. Moreover, a high number of false positives was obtained, indicating a low precision with the AD (Fig. 4a). To increase recall and precision of our detection method, we used both Decision Tree (DT) and random forest (RF) algorithms.

First, we used a random forest to analyze the identification profile of bNAbs with two classifying features and found that it allowed a clear decision boundary plot on the training dataset for bNAbs targeting the

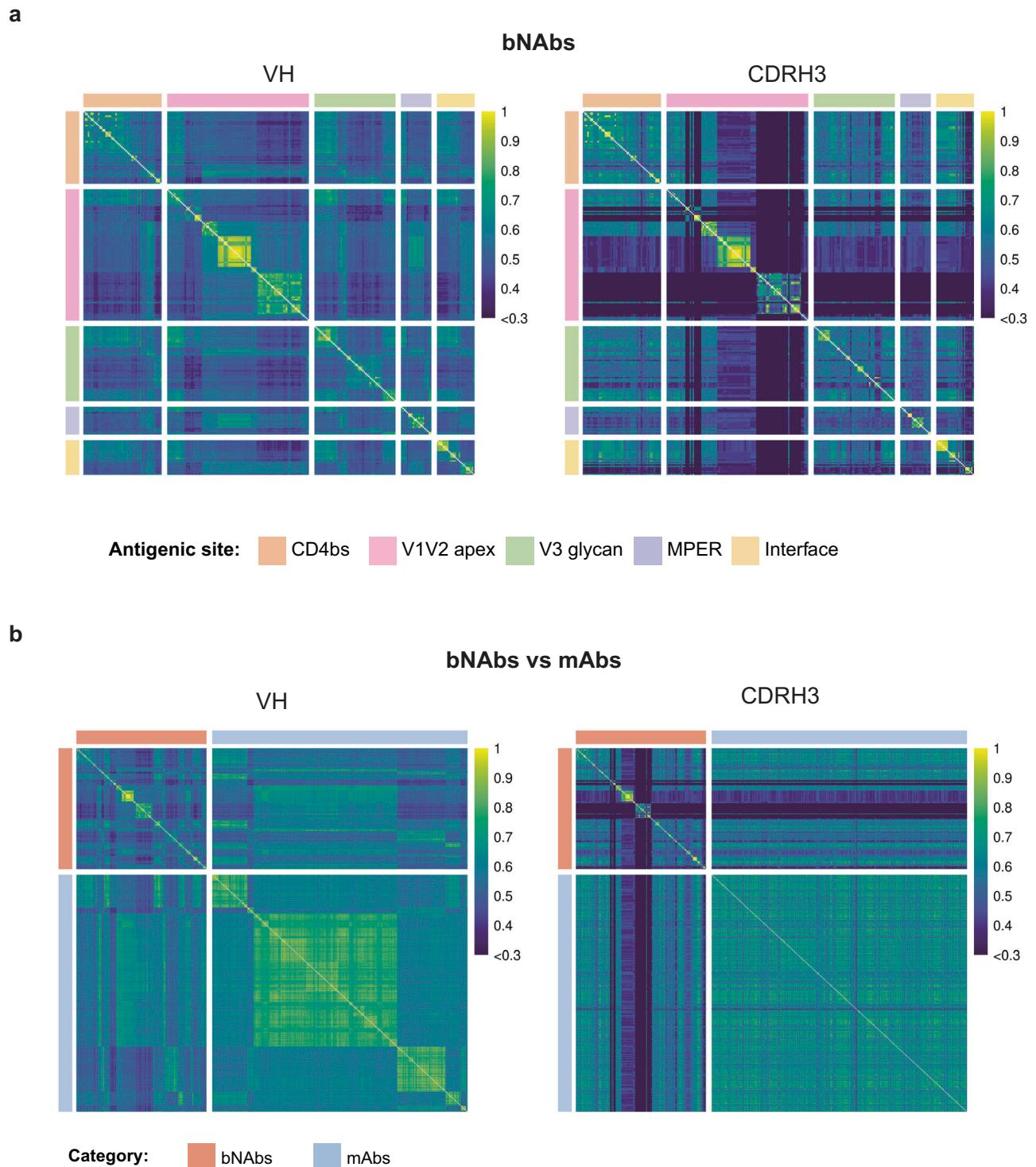


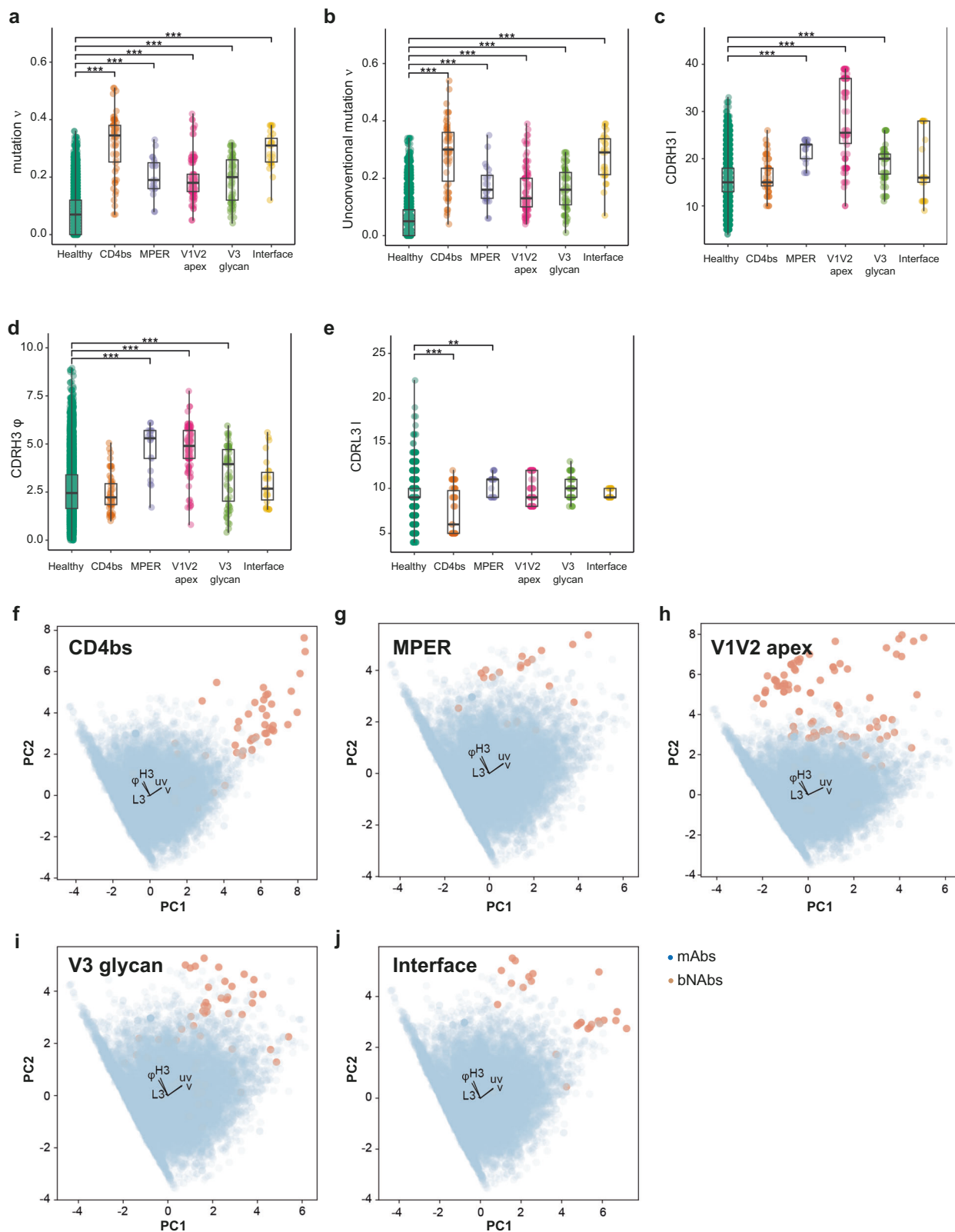
Fig. 2 | Sequence similarity matrices of HIV-1 bNAbs and control mAbs.
a Similarity matrices for 255 bNAbs grouped by antigenic site for the entire VH (left) or the CDRH3 only (right). **b** Similarity matrices of bNAbs versus mAbs with entire VH (left) and CDRH3 only (right). In the heatmaps, sequences are ranked based on their V and J genes. In both cases, matrices were created using ANARCI.

The similarity scores, ranging from 0 to 1, indicate the degree of similarity between sequences, with higher scores representing lower Levenshtein distances. **b** mAb sequences were downsampled to 500 to enable display. Source data are provided as a Source Data file.

interface or the V1V2 apex (Supplementary Fig. 3a, b). The receiver-operating characteristic (ROC) curve and corresponding AUC of 0.94 was obtained for V1V2 apex (Supplementary Fig. 3a) and 0.9 for interface (Supplementary Fig. 3b), indicating good classification performance for both antigenic sites. Furthermore, a measured AUC of 0.77 was obtained for bNAbs binding the CD4bs (Supplementary Fig. 3c).

However, the detection of bNAbs against other antigenic sites such as MPER (Supplementary Fig. 3c), and V3 loop (Supplementary Fig. 3e) was not satisfactory with an AUC close to 0.5 and 0.67, respectively.

Following this result, we allowed the DT and RF algorithms to use all available features, including VH and VL genes, and further optimized our models. We used the validation dataset to perform



hyperparameter tuning and systematically explore different combinations of hyperparameters. We based the classifiers' hyperparameter tuning on the false positives number, and for the hierarchical model of the DT, the cost complexity pruning parameter (α) was set to zero (Supplementary Fig. 4). Next, entropy was chosen as the quality measurement for the split in both DT and RF (further details are presented in "Methods"). The optimal parameters for our models enabled us to

achieve an overall very good performance, with a mean AUC of 0.87 (SD = 0.11) for the DT model and 0.95 (SD = 0.08) for the RF model (Supplementary Table 2). Notably, the mean precision score was very high, reaching 1 (SD = 0) for the RF model, while it was 0.5 (SD = 0.124) for the DT model. Finally, we used the test datasets and evaluated performance metrics, including AUC, precision, recall, and accuracy for the DT and RF models (Fig. 4a, c, d). The DT algorithm exhibited

Fig. 3 | Characteristics discriminating HIV-1 bNAbs from mAbs. Specific properties of antibodies that allow differentiation between bNAbs and mAbs depending on the antigenic site. Shown is boxplots with center line denoting the median value (50th percentile), while the black box contains the 25th to 75th percentiles of the dataset. The black whiskers mark the 5th and 95th percentiles. **a** Mutation frequency (v), **b** unconventional mutation frequency (uv), **c** CDRH3 length (H3), **d** CDRH3 hydrophobicity (ϕ), and **e** CDRL3 length (L3) were statistically compared with Kruskal–Wallis’s test followed by Dunn’s post hoc test. Only significant comparisons with mAbs are shown, with: * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.005$. **f–j** Principal component analysis (PCA) of the immunoglobulins using five features (v , uv , H3, ϕ , and L3). The feature weight for PC1 (Principal Component 1) and PC2

(Principal Component 2) is shown by black arrows. Each bNAb category is represented by a single plot per antigenic site, **f** CD4bs, **g** MPER, **h** VIV2 apex, **i** V3 glycan, and **j** gp120/gp41 interface. For data in **(a–e)**, sequences number is $n = 14,962$ for healthy, $n = 54$ for CD4bs, $n = 21$ for MPER, $n = 98$ for VIV2 apex, $n = 56$ for V3 glycan and $n = 26$ for interface. Adjusted P values with the Holm method are as follows: **(a)** healthy vs CD4bs $P = 2.04e-31$, MPER $P = 1.72e-10$, VIV2 $P = 5.94e-44$, V3 $P = 1.06e-20$ and interface $P = 3.00e-17$. **b** Healthy-CD4bs $P = 1.51e-30$, MPER $P = 4.87e-10$, VIV2 $P = 7.90e-34$, V3 $P = 3.08e-21$ and interface $P = 1.03e-16$. **c** Healthy-MPER $P = 1.04e-09$, VIV2 $P = 5.33e-45$ and V3 $P = 3.12e-10$. **d** Healthy-MPER $P = 3.97e-08$, VIV2 $P = 8.92e-36$ and V3 $P = 6.08e-04$. **e** Healthy-CD4bs $P = 8.89e-10$ and MPER $P = 0.02$. Source data are provided as a Source Data file.

superior recall and precision performance compared to the AD algorithm, while the RF algorithm demonstrated even higher performance, achieving a minimum AUC of 0.92 for all tested antigenic sites. It achieved a precision of 1 for almost all antigenic sites (0.83 for the interface). Moreover, an AUC of 1.0 and 0.95 for the MPER and interface site respectively, but also 0.95 for the V3 glycan, demonstrating that RF had the best performance as expected. Next, we reviewed the selected parameters used as RF classifiers. Interestingly, among the seven most important features, some were shared between the antigenic sites, while others were distinct. For instance, the mutation frequency and hydrophobicity of the CDRH3 were often key predictors (Fig. 4e). While the mutation frequency was an expected characteristic due to the long affinity maturation process required to obtain bNAbs, hydrophobicity of the CDRH3 might be interpreted as a consequence of the important glycan shield surrounding gp120/gp41 trimer. The frequency of unconventional mutations and length of the CDR3 light chain appears as an important feature for anti-CD4bs and is in agreement with reported bNAbs of this antigenic class⁴¹. The essential features associated with anti-interface bNAbs were also characterized by their mutation frequency both conventional and unconventional. The VIV2 apex binders were classified based on their CDRH3 lengths. Interestingly, bNAbs targeting the V3 glycan and MPER have a more balanced classification with features such as CDRH3 hydrophobicity, mutation, and CDRH3 length sharing similar weights (Fig. 4e). The immunoglobulin variable VH5-51 gene segment was associated with bNAbs targeting the V3 glycan as previously reported for 35% of human anti-V3 bNAbs⁴². As a final validation step, we compared the prediction results of each algorithm. These results indicate that predictors are specific to the antigenic site, even if the mutation frequency and hydrophobicity were always important.

Altogether, we observed that the different methods (AD, DT, and RF) identified the same true positives, while there was minimal overlap in false positives (Supplementary Fig. 5). To increase robustness and decrease the likelihood of false positive predictions, we combined different classifiers using the Super Learner Ensembles algorithm (SL) as an additional validation step⁴³. SL is an algorithm combining multiple models to make an “ensemble” prediction. The SL algorithm exhibited very high accuracy and precision performance with a score of 1 for all antigenic sites (Supplementary Fig. 6a) and achieved high performance for the MPER, VIV2 apex, and interface antigenic sites with a minimum AUC of 0.92 (Supplementary Fig. 6b). In contrast, the AUC was lower for the CD4bs, and V3 glycan antigenic sites (0.77 and 0.68), with a recall score of 0.53 and 0.35, respectively (Supplementary Fig. 6a). Based on the performance of our machine-learning approach for the Rapid Automatic Identification of bNAbs from Immune Repertoires (RAIN), we decided to use it on experimental samples in an effort to discover new bNAbs.

Experimental validation of the pipeline using de novo immune repertoires

To identify potential bNAbs, we investigated the neutralizing activity of purified immunoglobulin G (IgG) from the sera of different HIV-1 infected donors. Polyclonal IgGs from the serum of donors were

purified with protein G resin and tested on the global HIV-1 panel of reference strains, containing strains that are representative of the global epidemic^{44,45}. Interestingly, we observed that sera of donors 3, 11 and to some extent donor 9 had a broad neutralizing activity (Fig. 5a). In contrast, sera from donors 1, 2, 5, 6, 7, and 8 were able to neutralize only one or two viral strains (Fig. 5a). Based on this result, we selected the serum of donor 3 as test sample for the bNAb identification, while sera of donors 1 and 2 were selected as negative controls. We isolated IgG-class-switched B cells from peripheral blood mononuclear cells (PBMCs) of the different donors and performed single-cell sequencing of the B-cell receptors (BCRs) (B3, G3, S4, and G4). Importantly, no enrichment step was applied for B-cell sorting to ensure an unbiased repertoire for the downstream analysis. After filtering for error-corrected and productive sequences, we successfully reconstituted a set of 15,713 IgG sequences for donor 3. As a negative control, we sequenced BCRs from IgG+ memory B cells of donors 1 and 2 (that did not have sera with broad neutralization activity), which resulted in the acquisition of 8347 IgG sequences (D1 and D2). Interrogation of the RAIN pipeline on the sequences obtained from donor 3, led to the identification of several potential bNAbs, but only 3 were recognized by the three algorithms out of 15,713 paired sequences. To assess the specificity of RAIN on HIV samples, we decided to analyze B-cell repertoires from individuals exposed to a different viral infection or post-vaccination as an alternative control. We used sequences obtained from an Influenza vaccinated donor at days 7 and 9 post-vaccination⁴⁶. These sequences correspond to three sequencing runs of 4691, 8222, and 8052 paired BCRs sequences, respectively. While these repertoires contain Influenza bNAbs⁴⁶, our models did not detect any HIV bNAbs, indicating their specificity toward anti-HIV sequences (Supplementary Fig. 7a). To further confirm the three predicted HIV-1 bNAbs found in donor 3, we used the SL model, which identified thirteen potential bNAbs in this donor: six predicted to bind to the CD4-binding site, one to VIV2 apex, and six interface binders (Supplementary Fig. 7b). Interestingly, SL confirmed our predicted bNAbs, but also identified an anti-VIV2 apex binder in donor 2. These three potential bNAbs were constantly identified as CD4 binders (bNAb2101, bNAb4251, and bNAb1586) and belong to the VRC01 class of bNAbs (Supplementary Fig. 8).

Binding and neutralization properties of the identified bNAbs

To consolidate these findings, we cloned the three potential bNAbs and some additional antibodies as negative control (hereafter referred to mAbs). bNAbs and mAbs were recombinantly produced to test their specificity and neutralizing activities. We first assessed their binding to the envelope trimer SOSIP (using the clade A gp140 envelope stabilized prefusion trimer BG505 DS-SOSIP)^{47,48}, which is known to bind bNAbs that are representative of the majority of the known gp120 neutralizing antibody class^{49,50}. Using biolayer interferometry (BLI), we detected high-affinity interactions between all the identified bNAbs and SOSIP, characterized by an apparent equilibrium dissociation constant (K_D) of 115 ± 15 nM, 3 ± 0.6 nM, and 0.4 ± 0.03 nM and for bNAbs 1586, 2101, and 4251 respectively. In contrast, no interaction could be detected between the control mAbs and SOSIP (Fig. 5b and Supplementary Fig. 9a). To further characterize these interactions, we

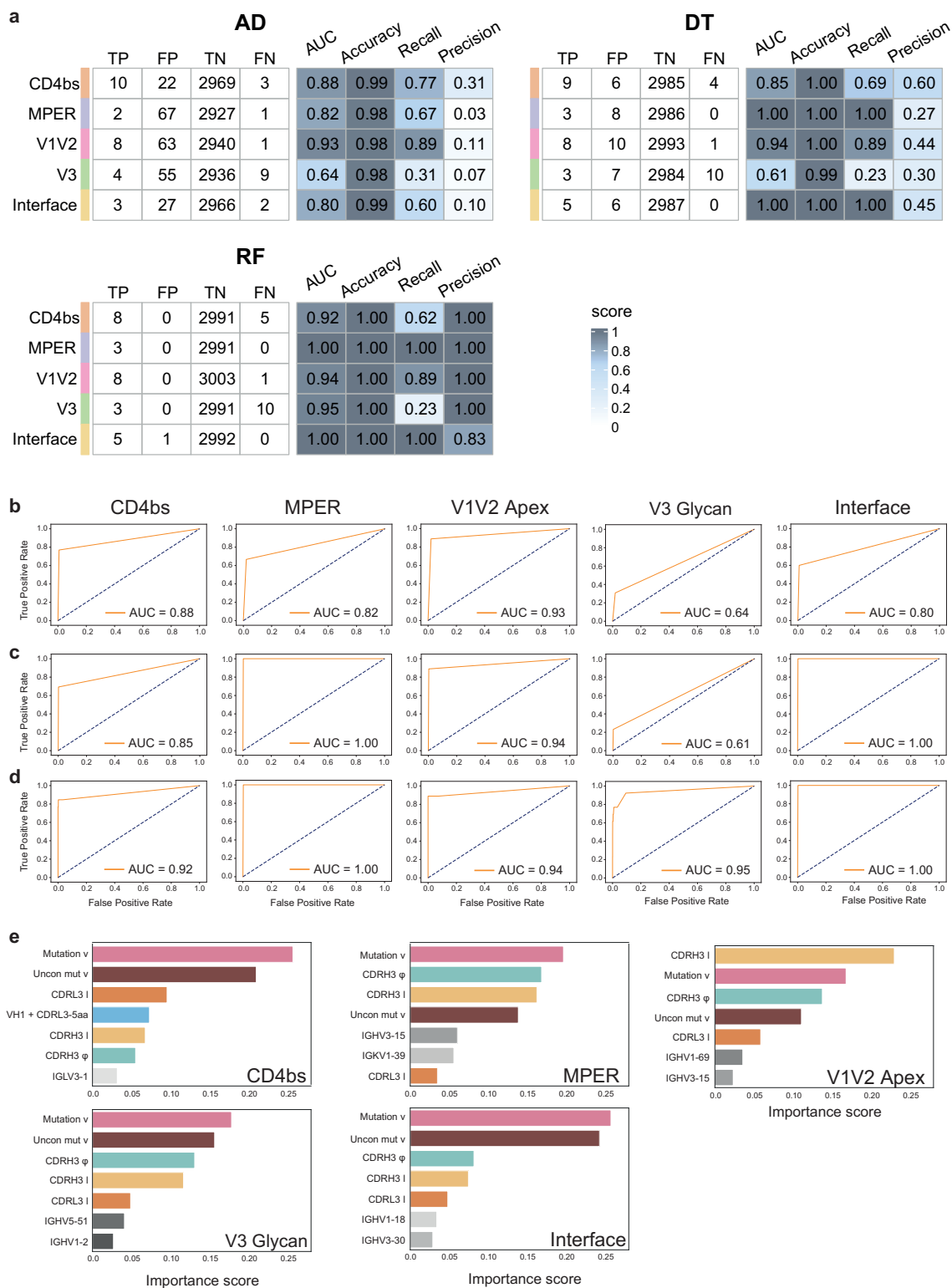
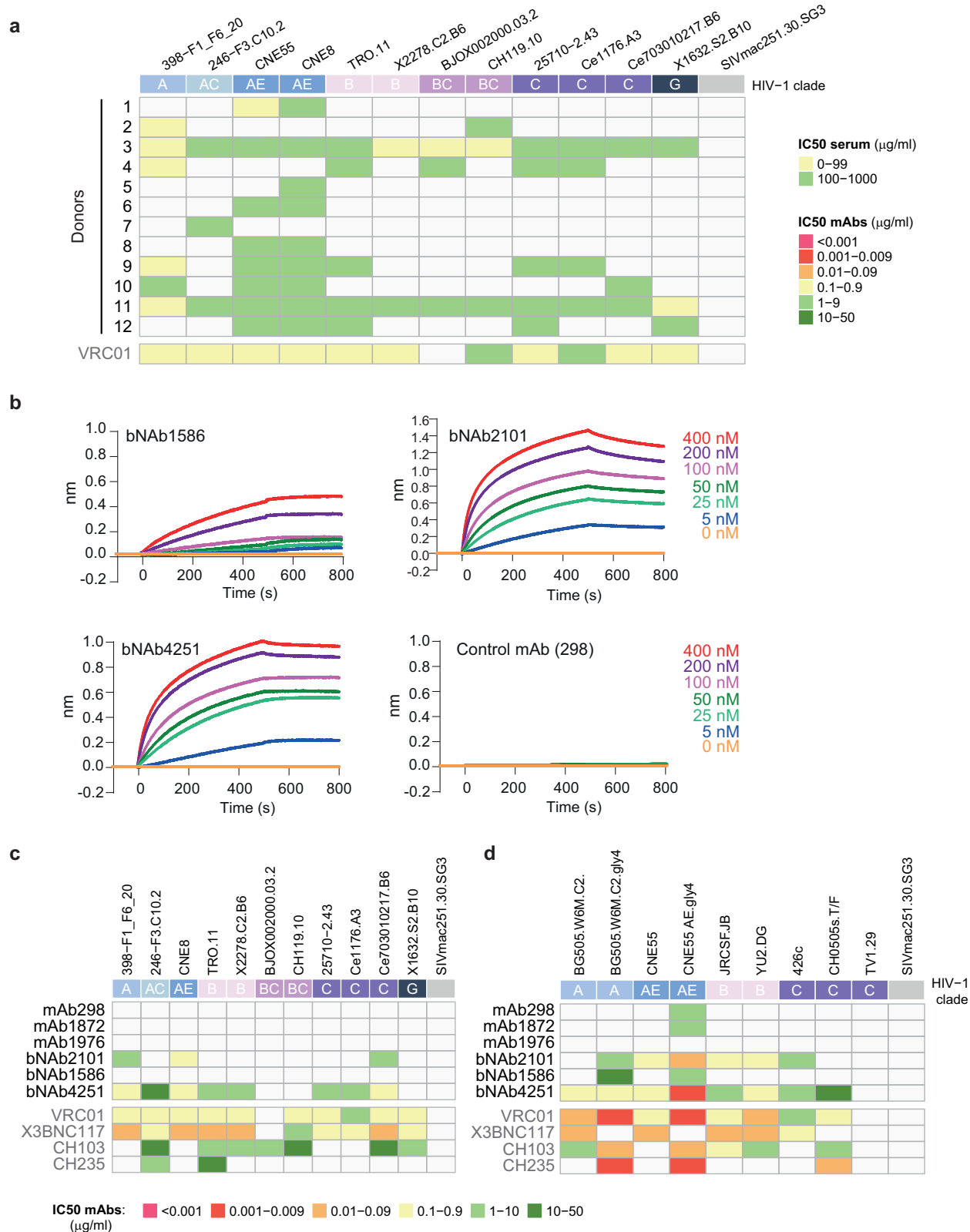


Fig. 4 | Performance of RAIN machine-learning models. **a** Performance metrics of the three algorithms using the test dataset with Accuracy = $(TP + TN)/(TP + FP + TN + FN)$, Recall = $TP/(TP + FN)$, and Precision = $TP/(TP + FP)$. **b–d** Receiver-operating characteristic (ROC) curves and corresponding area under the curve

(AUC) statistics for each bNAb antigenic site with the test dataset. Each row represents one algorithm, **b** AD, **c** DT, and **d** RF. **e** Most important features with their scores for each bNAb classified by binding antigenic site using the Random Forest classifier.

calculated the affinity of the fragment antigen binding (Fab) to SOSIP trimers and obtained a K_D 5 ± 2.4 nM, and 17.5 ± 4 nM for Fab4251 and Fab2101, respectively (Supplementary Fig. 9b). Of note, Fab1586 demonstrated poor affinity with a K_D measure of $1 \mu\text{M}$ (Supplementary Fig. 9b). To investigate the neutralization potency of these bNAbs, we

sought to determine their IC_{50} using the global HIV-1 panel strains on TZM-bl cells^{44,45}. We observed a broad neutralization activity across tiers and viral clade for bNAb4251, with a geometric mean IC_{50} of $1.8 \mu\text{g/ml}$ (Fig. 5c). Moreover, bNAb2101 could also neutralize different HIV strains and more specifically clade AE viruses, however, its



neutralization profile could not be considered broad (Fig. 5c). In contrast, bNAb1586 was a relatively poor neutralizer, only able to inhibit the CNE55 strain at 38 µg/ml (Fig. 5d). Importantly, none of the antibodies had an effect on the SIVmac251.30.SG3 virus indicating a specific neutralization activity. Overall, bNAb4251 could neutralize about 80% of the tested viruses but was not active against the TV1.29 and BJOX002000, similar to VRC01, which targets the CD4-binding site⁵¹.

Since both potential bNAbs were predicted to target the CD4-binding site, we further tested their neutralization potential on virus strains lacking the glycosylation surrounding the CD4bs such as the BG505.W6M.C2 strain with residues T332N (C2) or N197, N276, N363, and N462 (gly4) and other mutations previously described⁵² (Fig. 5d). Finally, clade C strains were also used, since the glycan at position 362 is naturally absent. The neutralization profile showed an increased

Fig. 5 | HIV Env binding and neutralization assays of serum and IgG samples. **a** Neutralization assays were performed against 12 viruses from clades A, AC, AE, B, BC, C, and G of tiers 2. The colors of the heatmap correspond to the IC₅₀ of the sera in micrograms per ml. The SIVmac251.30.SG3 virus is used as a negative control. **b** Antibody–SOSIP interactions were determined by biolayer interferometry (BLI). The mAbs or bNAbs were loaded on a protein A biosensor, dipped into a solution of the SOSIP trimer at different concentrations (ranging from 5 to 400 nM), and the nm shift was recorded. BLI sensorgrams are representative examples of

experiments repeated two times ($n > 2$). **c, d** Neutralization assays were performed against twelve viruses from clades A, AC, AE, B, BC, C, and G of tiers 2. **c** The colors of the heatmap correspond to the IC₅₀ in micrograms per ml, for each antibody. The SIVmac251.30.SG3 virus is used as a negative control. **d** Neutralization assays were performed against glycan-mutated viruses to support epitope mapping to the CD4-binding site. Neutralization assay experiments were repeated two times ($n > 2$). Source data are provided as a Source Data file.

potency specifically for the glycan mutations surrounding CD4bs, suggesting again that these antibodies target the CD4bs (Fig. 5d).

Binding mode of Fab4251 and Fab2101 to BG505 DS-SOSIP complex

Based on the affinity and neutralization potency of Fab2101 and Fab4251, we decided to investigate their binding mode using electron microscopy. We incubated BG505 DS-SOSIP with either 3 molar excesses of Fab2101 or Fab4251 and imaged the complex after 30 min of incubation at room temperature. We used single-particle negative stain electron microscopy (nsEM) to assess the sample purity and to map antibody epitopes on the viral glycoproteins⁵³ (Fig. 6a, b). Particles were picked from raw micrographs, stacks were created, followed by a reference-free 2D classification. SOSIP complexes appear as homogeneous trimers, as described previously⁵⁴. We identified that both Fabs bound to the soluble trimers in a manner similar to CD4bs-directed bNAbs, approaching the gp120 protomers from the side. To understand the molecular mechanism of the broad neutralization capacity by bNAb4251, we decided to perform cryo-EM of the Fab4251 in complex with the soluble native-like trimer BG505 DS-SOSIP⁵⁵. After several rounds of 2D and 3D classification (Supplementary Fig. 10), we could segregate SOSIP trimers with zero, or one Fab bound. We solved the structure of the complex at a resolution of 3.8 Å (Fig. 6c and Supplementary Table 3). As predicted by RAIN, Fab4251 interacts with the CD4bs of the trimer and makes multiple contacts with both heavy and light chains (Fig. 6c, d). In total, fifty-one residues of the Fab interact with fifty-six residues on gp120, to bury a surface area (bsa) of 950 Å². The interaction is principally dictated by the heavy chain with 700 Å² bsa, while the light chain buries 250 Å² of the gp120 surface (Fig. 6d). The CDRH2 makes most of the contact, totaling a bsa of 528 Å², a binding mode similar to the previously described interaction of the CD4 receptor with gp120 (Fig. 6g). The previously solved interaction of CD4 with gp120 revealed that two amino acids, F43 and N59 of CD4, make multiple contacts centered on residues N368, E370, and W427 of gp120^{56–58} (Fig. 6g). Interestingly, H54 of CDRH2 seems to mediate similar interactions with amino acids of the “F43 cavity” located at the interface between the inner and outer gp120 domains (Fig. 6g). Previously reported bNAbs targeting the CD4bs have been classified into two groups based on their mode of recognition, the VRC01 class (3BNC117, N6, N49P7, 3BNC60, VRC-PG20, NIH45-46, VRC-CH31, and 12A12) and the non-VRC01 classes (CH103, 8ANCI31, VRC13, and VRC16)⁵⁹. Structural investigations revealed that Fab4251 possesses an angle of approach similar to VRC01 (Fig. 6h), a result in agreement with its CDRH2-mediated contact on gp120, indicating that it belongs to the same antibody class (Fig. 6h). The light chain also participates in the interaction with the 5-residue LCDR3 QxxEx motif and a deletion in CDRL1 to accommodate the gp120 N276-glycan²⁸, a feature associated with VRC01-class antibodies.

Discussion

The advent of single-cell technologies resulted in the growing availability of paired full-length variable heavy and light-chain BCR sequences. Therefore, immune repertoire sequencing coupled to artificial intelligence holds great promise to improve diagnosis and treatment for numerous immune-related or infectious diseases⁶⁰. The identification of specific sequences involved in an immune response has already been

successfully used in research settings to elucidate the role of immune dysregulation in conditions such as systemic lupus erythematosus, rheumatoid arthritis, type 1 diabetes, multiple sclerosis, Grave’s disease, Crohn’s disease, and many others⁶¹. However, limitations exist and only a few studies examined the benefit of incorporating full-length variable regions from heavy and light-chain sequences to predict antibody specificity. Those studies are based on sequence-based embedding models^{62,63}. Other efforts have focused on finding amino acid sequence similarity to an already known antibody. The similarity approaches led to important scientific and medical discoveries^{64–66}, but hold some limitations when the sequences are very divergent.

In this study, we present RAIN, a pipeline based on two innovative technologies, single-cell BCR sequencing and machine learning to identify bNAbs against HIV-1, based on their binding site. Our approach differs from other methods as the parameters required for the identification are derived from selected characteristics, that are inferred from the amino acid sequences using Immcantation annotations. We demonstrate that five specific characteristics were sufficient to separate bNAbs from mAbs (non-bNAbs) into two distinct clusters within each category of antigenic sites. In addition, we identify the frequency of unconventional mutations as a key factor to define HIV-1 bNAbs. Former studies reported the presence of mutations in the frameworks of bNAbs and correlated them with the binding affinity to the CD4bs^{35,67}. Our results suggest that these mutations are important characteristics for all bNAbs. This can be interpreted as a consequence of the time needed for the maturation process of bNAbs or as a modification of the immune system in response to chronic infection.

Performing a PCA analysis across all five antigenic sites, we observed that despite their sequence divergences, the weights of the features exhibited striking similarities. This result could be interpreted as an additional level of immune escape that was not studied yet^{68,69}. The RAIN approach can achieve a precision of 1 for almost all antigenic sites and can be applied easily on any immune repertoire or already isolated antibody sequences to identify HIV-1 bNAbs. To our knowledge, this study pioneer in silico identification of specific antibodies, that could not have been identified by sequence alignment. Importantly, another distinct aspect of our work is the experimental validation with de novo data. Data were corroborated by functional cloning, expression and purification of the antibodies, and functional neutralization assays. Moreover, we characterized the bNAb4251 binding to DS-SOSIP at almost atomic resolution using cryo-EM. In summary, our approach offers an innovative, straightforward method to search and identify antibodies in immune repertoires, accelerate antibody discovery, and might shed light on potentially unexplored mechanisms of HIV-1 immune escape.

Methods

Ethics statement

The research complies with all relevant ethical regulations and informed consent was obtained by all study participants ($n = 25$, 16 females and 9 males). Study protocols were approved by the Ethikkommission beider Basel (EKBB; Basel, Switzerland; reference number 342/10), the Ifakara Health Institute Institutional Review Board (Reference number IHI/IRB/No. 24-2010), and the National Institute for Medical Research (NIMR; Dar es Salaam, United Republic of Tanzania; reference number NIMR/HQ/R.8a/Vol.IX/1162).

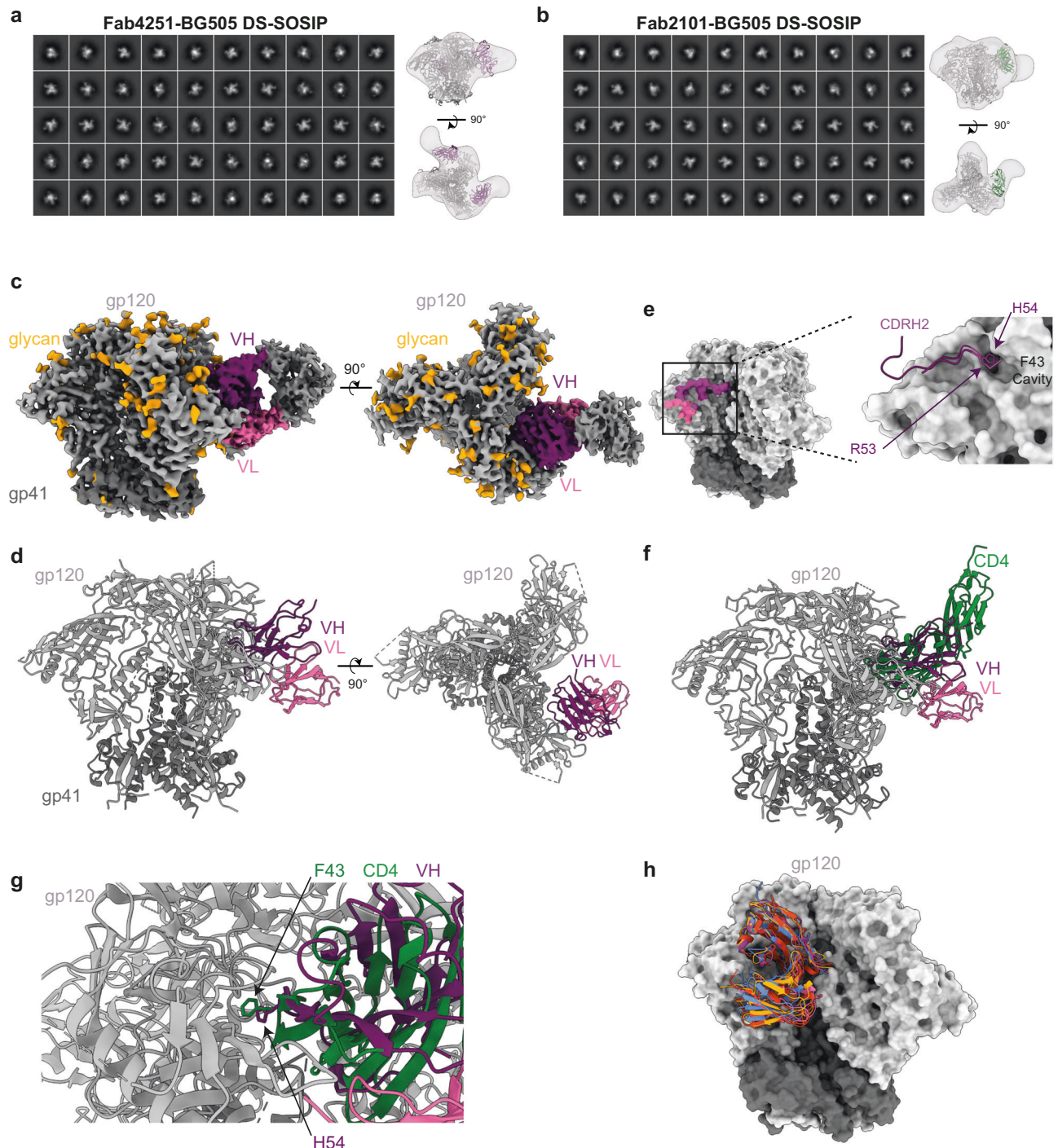


Fig. 6 | Fab4251 and Fab2101 interaction with BG505 DS-SOSIP. **a** 3D reconstruction of Fab4251-SOSIP complex by nEM. **b** 3D reconstruction of Fab2101-SOSIP complex by nEM. **c** Side and top views of the cryo-EM density map of the Fab4251-DS-SOSIP complex, with gp120 in light gray, gp41 in dark gray, VH in violet, and VL in pink. **d** Atomic model of Fab4251-DS-SOSIP complex shown in cartoon representation. **e** Footprint representation of the heavy and light-chain binding

surface on DS-SOSIP, colored according to **(c)**. Inlet on the right represents the CDRH2 loop in violet, with H54 in the F43 cavity. **f** Overlay of CD4 receptor (green) bound to SOSIP (PDB.5UIF) and Fab4251 (violet). **g** Close view of VH H54 from Fab4251 and F43 in CD4. **h** Overlay of VRC01-class antibodies on SOSIP with Fab4251 (violet), VRC01 (PDB.6V8X, green), PG04 (PDB.4I3S, red), and 3BNC60 (PDB.4GW4, orange).

Serum IgG isolation

Serum samples from HIV-1-infected individuals were incubated with Protein G Sepharose (GE Life Sciences) 4 °C for 1 h. IgGs were eluted from chromatography columns using 0.1M glycine (pH = 2.9) into 0.1M Tris (pH = 8.0)⁷⁰. Samples were run through Zeba Spin Desalting Columns 7 K MWCO (Thermo Scientific, 89882) Concentrations of

purified IgGs were determined by UV/Vis spectroscopy (A280) on a Nanodrop 2000 and samples were stored at -80 °C.

B-cell sorting

The CD19+ cell fraction was enriched from PBMCs by positive selection with CD19 magnetic microbeads (Miltenyi Biotech) and subsequently

stained on ice for 30 min with the following fluorochrome-labeled mouse monoclonal antibodies: CD20-PE-Cy7 (dilution 1:50, clone L27, catalog no. 335793, BD Biosciences) and F(ab')₂-Goat anti-Human IgG Fc secondary antibody, APC (dilution 1:100, RRID: AB_2337695, Jackson ImmunoResearch). Cells were sorted to over 98% purity on a FACS Aria III (BD) using the following gating strategy: circulating memory B cells were sorted as CD20+ IgG+ cells. FACS-sorted cells were collected in 6 μ l FCS in Eppendorf tubes that were pre-coated overnight with 2% BSA.

Single-cell BCR-seq library preparation and sequencing

10X Genomics. The 5' single-cell VDJ libraries were generated using Chromium Next GEM Single Cell V(D)J Reagent kit v.1, 1.1 or v.2 (10X Genomics) according to the manufacturer's protocol. Paired heavy and light-chain BCR libraries were prepared from the sorted B-cell populations. Briefly, up to 20,000 memory B cells per well of 10X chip were loaded in the 10X Genomics Chromium Controller to generate single-cell gel beads in emulsion. After reverse transcription, gel beads in the emulsion were disrupted. Barcoded complementary DNA was isolated and used for the preparation of BCR libraries. All the steps were followed as per the manufacturer's instructions in the user guide recommended for 10X Genomics kit v.1, 1.1, or 2. The purified libraries from each time point were pooled separately and sequenced on the NextSeq550 (Illumina) as per the instructions provided in 10X Genomics user guide for the read length and depth.

BD rhapsody. Memory B cells were targeted for single-cell targeted RNA-seq and BCR-Seq analysis using the BD Rhapsody Single-Cell Analysis System⁷¹ (BD Biosciences). Briefly, the single-cell suspension was loaded into a BD Rhapsody cartridge with >200,000 microwells, and single-cell capture was achieved by random distribution and gravity precipitation. Next, the bead library was loaded into the microwell cartridge to saturation so that the bead was paired with a cell in a microwell. The cells were lysed in a microwell cartridge to hybridize mRNA molecules onto barcoded capture oligos on the beads. These beads were then retrieved from the microwell cartridge into a single tube for subsequent cDNA synthesis, exonuclease I digestion, and multiplex-PCR-based library construction. Sequencing was performed on NovaSeq paired-end mode.

Singleron. Single-cell suspensions with 1×10^5 cells/mL in PBS were prepared. Then, the suspensions were loaded onto microfluidic devices, and scRNA-seq libraries were constructed according to the Singleron GEXSCOPE protocol in the GEXSCOPE Single-Cell RNA Library Kit (Singleron Biotechnologies)⁷². Individual libraries were diluted to 4 nM and pooled for sequencing. Pools were sequenced on an Illumina HiSeq X with 150 bp paired-end reads.

Recombinant antibody production

Expi293 cells (Thermo Fisher Cat No. A14527) were diluted to a final volume of 0.5 L at a concentration of 2.5×10^6 cells mL⁻¹ in Expi293 media⁷³. Heavy-chain and light-chain plasmids were complexed with Polyethyleneimine (Thermo Fisher) and added to the cells. On day five, cells were cleared from cell culture media by centrifugation at 10,000 \times g for 30 min, and the supernatant was subsequently passed through a 0.45- μ m filter. The supernatant containing the recombinant antibody was purified with the HiTrap Protein A HP column (Cytiva, 17040301) on the Äkta pure system (Cytiva). The resin was washed with 75 mL of phosphate-buffered saline (PBS). A total of 25 mL of 0.1 M glycine pH 2.9 were used to elute the antibody from the protein A resin. The acidic pH of the eluted antibody solution was increased to ~7 by the addition of 1 M Tris pH 8.0. The antibody solution was buffer exchanged to PBS by the HiPrep 26/10 Desalting column (GE Healthcare) or Size Exclusion Chromatography Superdex 16/600 HiLoad (Cytiva), filtered, snap-frozen in liquid nitrogen, and stored at -80 °C.

Fragment antigen binding (Fab) generation

For the Fab production, the heavy chain was engineered with a two amino acids glycine serine linker followed by a six-histidine tag and stop codon. Light and mutated heavy chains were transfected as described in the previous section. Cell supernatant was harvested five days post-transfection and purified by IMAC chromatography (HisTrap excel, Cytiva) using the elution buffer 25 mM Tris pH 7.4, 150 mM NaCl, 500 mM imidazole. The eluate was buffer exchanged to 25 mM Tris pH 7.4, 150 mM NaCl, 0.085 mM n-dodecyl β -D-maltoside (DDM) on a HiPrep 26/10 Desalting column (GE Healthcare), followed by Size Exclusion Chromatography on a Superdex 16/600 HiLoad column (Cytiva)⁷⁴. The sample was concentrated using an Amicon filter 10 kDa cutoff, snap-frozen, and stored at -80 °C until further use.

Recombinant HIV-1 envelope SOSIP gp140 production

BG505 DS-SOSIP trimer⁷⁵ production and purification were performed as previously described⁴⁸. Briefly, prefusion-stabilized Env trimer derived from the clade A BG505 strain was stably transfected in CHO-DG44 cells and expressed in ActiCHO P medium with ActiCHO Feed A and B as feed (Cytiva). Cell supernatant was collected by filtration through a Clarisolve 20MS depth filter followed by a Millistak + FOHC filter (Millipore Sigma) at 60 LMH. Tangential Flow Filtration was used to concentrate and buffer exchange clarified supernatant in 20 mM MES, 25 mM NaCl, pH 6.5. The trimer was then purified by ion exchange chromatography as described⁴⁸. Fractions containing the BG505 DS-SOSIP protein were pooled, sterile-filtered, snap-frozen, and stored at -80 °C.

IgG neutralization assay

Neutralization assays with IgGs against the 12-strain "global" virus panel, were performed in 96-well plates as previously described^{44,76,77}. Briefly, 293T-derived HIV-1 Env-pseudotyped virus stocks were generated by cotransfection of an Env expression plasmid and a pSG3 Δ Env backbone. Animal sera were heat-inactivated at 56 °C for 1 h and assessed at 8-point fourfold dilutions starting at 1:20 dilutions. Monoclonal antibodies were tested at 8-point fivefold dilutions starting at 50 μ g/ml or 500 μ g/ml. Virus stocks and antibodies (or sera) were mixed in a total volume of 50 μ l and incubated at 37 °C for 1 h. TZM-bl cells (20 μ l, 0.5 million/ml) were then added to the mixture and incubated at 37 °C. Cells were fed with 130 μ l cDMEM on day 2, lysed, and assessed for luciferase activity (RLU) on day 3. A nonlinear regression curve was fitted using the 5-parameter hill slope equation. The 50% and 80% inhibitory dilutions (ID₅₀ and ID₈₀) were determined for sera and the 50% and 80% inhibitory concentrations (IC₅₀ and IC₈₀) were determined for mAbs. All samples were tested in duplicates.

Biolayer interferometry

The biolayer interferometry experiments using SOSIP were performed as follows. All experiments were performed in reaction buffer (TBS pH 7.4 + 0.01% (w/v) BSA + 0.002% (v/v) Tween 20) at 30 °C using an Octet K2 instrument (ForteBio). Protein A (ForteBio) biosensor probes were first equilibrated in reaction buffer for 600 s. IgGs were diluted to 5 μ g/ml in reaction buffer and immobilized onto the protein A probes for 300 s, followed by a wash for 300 s in reaction buffer. The binding of SOSIP trimers to the IgGs was then measured at various concentrations for 500 s, followed by dissociation for 300 s in reaction buffer. Analysis was performed using the Octet software with bivalent analyte fitting for antibody binding and 1:1 analyte fitting for the interaction with Fabs. Association and dissociation curves are visualized by GraphPad Prism version 9.0.

Negative stain electron microscopy

The samples were adsorbed to a glow-discharged carbon-coated copper grid 400 mesh (EMS, Hatfield, PA, USA), washed with deionized water, and stained with a 1% uranyl acetate solution for 20 s.

Observations were made using an F20 electron microscope (Thermo Fisher, Hillsboro, USA) operated at 200 kV⁷³. Digital images were collected using a direct detector camera Falcon III (Thermo Fisher, Hillsboro, USA) 4098 × 4098 pixels. Automatic data collection was performed using the EPU software (Thermo Fisher, Hillsboro, USA) at a nominal magnification of ×62,000, corresponding to a pixel size of 1.65 Å using a defocus range from −1 μm to −2.5 μm. Image pre-processing, two-dimensional classification, and three-dimensional processing was done using the CryoSPARC software (Version 4.4)⁷⁸.

Cryo-EM sample preparation

BG505 DS-SOSIP trimers complexes were prepared using a stock solution of 5 mg/ml trimer incubated with a threefold molar excess of bNAb4251 for 10 min. To prevent aggregation and interaction of the trimer complexes with the air-water interface during vitrification, the samples were incubated in 25 mM Tris pH 7.4, 150 mM NaCl, 0.085 mM DDM. Samples were applied to plasma-cleaned QUANTIFOIL holey carbon grids (EMS, R2/2 Cu 300 mesh). The grid was plunge frozen using a Vitrobot MarkIV (Thermo Fisher, Hillsboro, USA) with humidity and temperature control.

Cryo-EM data collection

Grids were screened for particle presence and ice quality on a TFS Glacios microscope (200 kV), and the best grids were transferred to a TFS Titan Krios G4. Cryo-EM data were collected using a TFS Titan Krios G4 transmission electron microscope, equipped with a Cold-FEG on a Falcon IV detector in electron counting mode. Falcon IV gain references were collected just before data collection. Data were collected using TFS EPU v2.12.1 utilizing the aberration-free image shift protocol, recording four micrographs per ice hole. Movies were recorded at a magnification of ×165,000, corresponding to the 0.73 Å pixel size at the specimen level, with defocus values ranging from −0.9 to −2.4 μm. Exposures were obtained with 39.89 e[−] Å^{−2} total dose, resulting in an exposure time of ~2.75 s per movie. In total, 15,163 micrographs in EER format were collected.

Cryo-EM data processing and structure fitting

Data processing was performed with cryoSPARC (Version 4.4) including Motion correction and CTF determination⁷⁸. Particle picking and extraction (extraction box size 350 pixels²) were carried out using cryoSPARC Version 4.4⁷⁸. Next, several rounds of reference-free 2D classification were performed to remove artifacts and selected particles were used for ab initio reconstruction and hetero-refinement. After hetero-refinement, 72,497 particles contributed to an initial 3D reconstruction of 3.8 Å resolution (Fourier-shell coefficient (FSC) 0.143) with C1 symmetry. A model of a SOSIP trimer (PDB ID 4TVP)⁷⁹ or AlphaFold2 (ColabFold implementation) models of the 4251 Fab were fitted into the cryo-EM maps with UCSF ChimeraX (Version 1.5). These docked models were extended and rebuilt manually with refinement using Coot (Version 0.9.8.8) and Phenix (Version 1.21)^{80,81}. Figures were prepared in UCSF ChimeraX, and Pymol (Version 4.6)⁸². The numbering of Fab4251 is based on the Kabat numbering of immunoglobulin models⁸³. Buried surface area measurements were calculated within ChimeraX and PISA⁸⁴.

CATNAP sequences

For all antigenic sites, paired bNAb sequences were collected from the CATNAP database³² as of January 1, 2022 as nucleotide and amino acid sequences. First, the 249 heavy-chain and 240 light-chain nucleotide sequences were annotated with Igbblastn³⁶. Sequences were then processed and analyzed using the Immcantation Framework (<http://immcantation.org>) with MakeDB.py from Change-O v1.2.0 (with the options `--extended` `--partial`). Next, bNAb sequences were filtered by a dedicated Java script to keep only sequences with an annotated CDR3 and paired

sequences (VH + VK/L). Each paired antibody was associated with its targeting Env antigenic site, information provided by the database CATNAP text file (abs.txt as of January 1, 2022). The 27 CATNAP antibodies with only the protein sequences available were annotated with Igbblastp followed by MakeDB.py from Change-O v1.2.0 (with the options `igblast-aa` `--extended`). In parallel, using the fasta protein sequences, ANARCI⁸⁵ was used to identify the junction region. As for nucleotide sequences, paired and annotated CDR3 bNAb sequences were filtered in. In total, 255 bNAb sequences were collected. Repartition of the antigenic site is as follows: 54 bNAb target the CD4bs, 21 MPER, 98 VIV2, 56 V3, and 26 interface.

Paired B-cell receptor repertoires

For the training and evaluation of the machine-learning models, paired BCR repertoires of ten healthy donors were collected. The repertoires were obtained from various sources (Supplementary Data Files 1) and sequenced using 10X genomics technology. Annotation and processing of the sequences were done as previously described³⁹ and resulted in the generation of a customized AIRR format table containing 14,962 paired BCRs. For HIV-1 immune donors three different sequencing technologies were employed: 10X genomics (D1, D2, G3, and G4), Singleron (S4), and BD Rhapsody (B3). Single-cell sequencing of selected HIV-1 immune donors using Singleron technology was processed using telescope v1.14.1 (<https://github.com/singleron-RD/CeleScope>) with “flv_CR” mode utilizing cellranger v7.0.1. BD rhapsody single-cell sequencing was first processed using BD Rhapsody Targeted mRNA Analysis Pipeline (version 1.11) and then, using a custom script, the generated “VDJ_Dominant_Contigs.csv” file was converted into cellranger-like output files, namely filtered_contig_annotations.csv and filtered_contig.fasta. Lastly, the 10X Genomics single-cell sequencing was processed with cellranger v7.0.1. The cellranger output files of the different HIV-1 repertoires enabled us to annotate and process them as described earlier, resulting in a table of paired BCRs with AIRR characteristics. The six different experiments resulted in 2152 BCRs for D1, 6195 BCRs for D2, 4008 BCRs for B3, 3794 BCRs for G3, 3112 BCRs for S4, and 4799 BCRs for G4.

Sequence similarity matrices

All mAbs and bNAb VDJ protein sequences were initially aligned using ANARCI with IMGT format. Subsequently, employing a custom R script, two similarity matrices were generated: one encompassing the entire VDJ sequence (VH) and the other focusing solely on the CDRH3 region. For each pair of sequences, a Levenshtein distance was computed, yielding a similarity score ranging from 0 to 1 (higher score representing lower Levenshtein distance). Heatmaps were constructed with the pheatmap R package, to visualize the following comparisons: all five antigenic site categories of bNAb and the comparison bNAb versus mAb (mAb sequences were downsampled to 500 sequences). Sequences were ranked based on their V and J genes.

Data preprocessing

Using a custom script, AIRR characteristics were converted into our features of interest. The “mutation frequency” was calculated using the difference of residues between the protein sequence of the BCR and its germline sequence in the FWR1 + CDR1 + FWR2 + CDR2 + FWR3 regions (VH gene). The “framework mutation frequency” was calculated similarly but using only FWR1 + FWR2 + FWR3. The “hydrophobicity” of the CDRH3 sequences was computed using a customized score, with aromatic residues having the highest value (1 for W, 0.75 for Y, and 0.5 for F). Residues A, L, I, M, P, and V were set to 0.1, while the rest of the residues were set to zero. The values of all residues were summed up for each CDRH3. In addition, the length of the CDRH3, CDRL3, VH, and VL/K genes were considered

as features. Two extra features were added to be used by the anomaly detection algorithm: “VH1 + CDRL3 length of five residues” with a zero or one value designed for the bNAbs targeting the CD4bs and “VH1-69 + VK3-20 + GW motif in the CDRH3” with a zero or one value for the bNAbs targeting MPER.

Training and evaluation of machine-learning models

Three ML-based approaches were trained on the features table generated using BCRs obtained from healthy donors and bNAbs datasets, using Python v3.8.16 and scikit-learn v1.0.2. These algorithms were: Anomaly Detection (AD), Decision Tree (DT), and Random Forest (RF). For each antigenic site, the dataset was partitioned into training, validation, and test sets with a 60:20:20 ratio, setting random.seed to 1 for all models. For the AD model, bNAbs data were removed from the training set, since this algorithm only trains with non-anomaly data. For this model, the features with discrete values were first normalized using the preprocessing.normalize method (axis=0) from the scikit-learn library. Features exhibiting significantly different values from the normal distribution were selected for each antigenic site, which included the frequency of mutations in the V genes and in the frameworks. For CD4bs, we added the combined feature VH1 + CDR3L with a length of five residues. For MPER, we included the combined feature VH1-69, VK3-20, and the GW motif in CDRH3. In addition, CDRH3 hydrophobicity was added for MPER, VIV2, and V3. Lastly, CDRH3 length was incorporated for VIV2 and V3. Using the validation test, a multivariate normal random variable was calculated with the multivariate_normal function from the scipy package v1.8.0 and used for setting the optimal Epsilon parameter (ϵ) minimizing the false positive numbers. The Epsilon value was set to 619.55 for CD4bs, 231501.41 for MPER, 866803.64 for VIV2, 845445.99 for V3, and 24.36 for interface. Those threshold values were used on the test set to predict a BCR as an anomaly (bNAb) or not. For DT and RF models, V genes (for heavy and light chains) were one-hot encoded as a preprocessing step, resulting in a total of 122 features in the features table. Hyperparameter tuning was conducted using the validation dataset, minimizing the number of false positives. DT models were trained with a balanced class weight, the Entropy criterion for measuring the quality of splits, and the cost complexity pruning parameter alpha of zero. RF models were trained with 100 estimators, a balanced class weight, the Entropy criterion for measuring the quality of splits, maximum samples were set to 1.0, maximum depth of tree of “none”, maximum features of 11 ($\sqrt{122}$), and bootstrapping to build trees. Matplotlib library v3.6.2 was used to generate ROC plots from performance results and to generate the Venn diagrams showing the intersection of the number of true positives or false positives between the three models. The Super Learner Ensembles algorithm was implemented using the ML-Ensemble (mlens) v0.2.3 library. For each antigenic site, the dataset was partitioned into train and test sets with a 75:25 ratio. The Super Learner was created with the precision score as scorer parameter, a k-fold cross-validation of ten folds, and the option shuffle set to true. The following classifiers were used as based models in the Super Learner algorithm: DecisionTreeClassifier, SVC (Support Vector Classification), KNeighborsClassifier, AdaBoostClassifier, BaggingClassifier, RandomForestClassifier, and ExtraTreesClassifier. A LogisticRegression was used as the meta-model, with the solver parameter set to “lbfgs”.

Statistical analysis

Flow cytometric data were acquired using BD FACSDiva (v.9.0) software. Flow cytometric data were analyzed using FlowJo (v.10.7.1). Statistics were conducted using R Statistical Software (v4.2.1) and ggstatsplot package⁸⁶. The Complex Heatmap package was used for visualization⁸⁷. No statistical methods were used to predetermine the sample size. The experiments were not randomized, and investigators

were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data files for single-cell VDJ sequencing generated in this study have been deposited in the GEO database: [GSE229123](https://doi.org/10.1038/s41467-024-49676-1). Cryo-EM map generated in this study have been deposited on EMDB: [EMD-19665](https://doi.org/10.1038/s41467-024-49676-1), with PDB accession number [8S2E](https://doi.org/10.1038/s41467-024-49676-1). All other data supporting the findings of this study are available from the corresponding author on request. Source data are provided with this paper.

Code availability

The complete workflow and associated scripts are available on <https://github.com/MathildeFogPerez/manuscript-bnab-foglierini>. A set of instructions on how to use the workflow and completely reproduce the results shown herein is available there.

References

- Landovitz, R. J., Scott, H. & Deeks, S. G. Prevention, treatment and cure of HIV infection. *Nat. Rev. Microbiol.* **21**, 657–670 (2023).
- Haynes, B. F. & Burton, D. R. Developing an HIV vaccine. *Science* **355**, 1129–1130 (2017).
- Sok, D. & Burton, D. R. Recent progress in broadly neutralizing antibodies to HIV. *Nat. Immunol.* **19**, 1179–1188 (2018).
- Bailey, J., Blankson, J. N., Wind-Rotolo, M. & Siliciano, R. F. Mechanisms of HIV-1 escape from immune responses and anti-retroviral drugs. *Curr. Opin. Immunol.* **16**, 470–476 (2004).
- Malim, M. H. & Emerman, M. HIV-1 sequence variation: drift, shift, and attenuation. *Cell* **104**, 469–472 (2001).
- Liao, H. X. et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).
- Zhou, T. & Xu, K. Structural features of broadly neutralizing antibodies and rational design of vaccine. *Adv. Exp. Med. Biol.* **1075**, 73–95 (2018).
- Roskin, K. M. et al. Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nat. Immunol.* **21**, 199–209 (2020).
- Pantaleo, G., Correia, B., Fenwick, C., Joo, V. S. & Perez, L. Antibodies to combat viral infections: development strategies and progress. *Nat. Rev. Drug Discov.* **21**, 676–696 (2022).
- Shingai, M. et al. Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277–280 (2013).
- Barouch, D. H. et al. Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224–228 (2013).
- Parsons, M. S. et al. Partial efficacy of a broadly neutralizing antibody against cell-associated SHIV infection. *Sci. Transl. Med.* **9**, eaaf1483 (2017).
- Gautam, R. et al. A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges. *Nature* **533**, 105–109 (2016).
- Halper-Stromberg, A. et al. Broadly neutralizing antibodies and viral inducers decrease rebound from HIV-1 latent reservoirs in humanized mice. *Cell* **158**, 989–999 (2014).
- Caskey, M., Klein, F. & Nussenzweig, M. C. Broadly neutralizing anti-HIV-1 monoclonal antibodies in the clinic. *Nat. Med.* **25**, 547–553 (2019).
- Mendoza, P. et al. Combination therapy with anti-HIV-1 antibodies maintains viral suppression. *Nature* **561**, 479–484 (2018).
- Gaebler, C. et al. Prolonged viral suppression with anti-HIV-1 antibody therapy. *Nature* **606**, 368–374 (2022).

18. McCoy, L. E. The expanding array of HIV broadly neutralizing antibodies. *Retrovirology* **15**, 70 (2018).
19. Krebs, S. J. et al. Longitudinal analysis reveals early development of three MPER-directed neutralizing antibody lineages from an HIV-1 infected individual. *Immunity* **50**, 677–691.e613 (2019).
20. Schriek, A. I., Aldon, Y. L. T., van Gils, M. J. & de Taeye, S. W. Next-generation bNAbs for HIV-1 cure strategies. *Antivir. Res.* **222**, 105788 (2023).
21. Mahomed, S., Garrett, N., Baxter, C., Abdool Karim, Q. & Abdool Karim, S. S. Clinical trials of broadly neutralizing monoclonal antibodies for human immunodeficiency virus prevention: a review. *J. Infect. Dis.* **223**, 370–380 (2021).
22. Sneller, M. C. et al. Combination anti-HIV antibodies provide sustained virological suppression. *Nature* **606**, 375–381 (2022).
23. Karuna, S. T. & Corey, L. Broadly neutralizing antibodies for HIV prevention. *Annu Rev. Med.* **71**, 329–346 (2020).
24. Marks, C. & Deane, C. M. How repertoire data are changing antibody science. *J. Biol. Chem.* **295**, 9823–9837 (2020).
25. Kim, J., McFee, M., Fang, Q., Abdin, O. & Kim, P. M. Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol. Sci.* **44**, 175–189 (2023).
26. Akbar, R. et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* **14**, 2008790 (2022).
27. Scheid, J. F. et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633–1637 (2011).
28. West, A. P. Jr., Diskin, R., Nussenzweig, M. C. & Bjorkman, P. J. Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proc. Natl. Acad. Sci. USA* **109**, E2083–E2090 (2012).
29. Jardine, J. G. et al. HIV-1 VACCINES. Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen. *Science* **349**, 156–161 (2015).
30. Liao, H. et al. Contribution of V(H) replacement products to the generation of anti-HIV antibodies. *Clin. Immunol.* **146**, 46–55 (2013).
31. Willis, J. R. et al. Human immunoglobulin repertoire analysis guides design of vaccine priming immunogens targeting HIV V2-apex broadly neutralizing antibody precursors. *Immunity* **55**, 2149–2167.e2149 (2022).
32. Yoon, H. et al. CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res.* **43**, W213–W219 (2015).
33. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).
34. Shen, C. H. et al. VRC34-antibody lineage development reveals how a required rare mutation shapes the maturation of a broad HIV-neutralizing lineage. *Cell Host Microbe* **27**, 531–543.e536 (2020).
35. Wiehe, K. et al. Functional relevance of improbable antibody mutations for HIV broadly neutralizing antibody development. *Cell Host Microbe* **23**, 759–765.e756 (2018).
36. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
37. Nouri, N. & Kleinstein, S. H. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* **34**, i341–i349 (2018).
38. Foglierini, M., Pappas, L., Lanzavecchia, A., Corti, D. & Perez, L. AncesTree: an interactive immunoglobulin lineage tree visualizer. *PLoS Comput. Biol.* **16**, e1007731 (2020).
39. Phad, G. E. et al. Clonal structure, stability and dynamics of human memory B cells and circulating plasmablasts. *Nat. Immunol.* **23**, 1076–1085 (2022).
40. Steinwart, I., Hush, D. & Scovel, C. A classification framework for anomaly detection. *J. Mach. Learn. Res.* **6**, 211–232 (2005).
41. Zhou, T. et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258 (2013).
42. Gorny, M. K. et al. Preferential use of the VH5-51 gene segment by the human immune response to code for antibodies against the V3 domain of HIV-1. *Mol. Immunol.* **46**, 917–926 (2009).
43. van der Laan M. J., Polley E. C., Hubbard A. E. Super Learner. *Stat. Appl. Genet. Mol. Biol.* <https://doi.org/10.2202/1544-6115.1309> (2007).
44. deCamp, A. et al. Global panel of HIV-1 Env reference strains for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **88**, 2489–2507 (2014).
45. Schommers, P. et al. Restriction of HIV-1 escape by a highly broad and potent neutralizing antibody. *Cell* **180**, 471–489.e422 (2020).
46. Horns, F., Dekker, C. L. & Quake, S. R. Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. *Cell Rep.* **30**, 905–913.e906 (2020).
47. Chuang, G.-Y. et al. Structure-based design of a soluble prefusion-closed HIV-1 Env trimer with reduced CD4 affinity and improved immunogenicity. *J. Virol.* **91**, e02268–16 (2017).
48. Gulla, K. et al. A non-affinity purification process for GMP production of prefusion-closed HIV-1 envelope trimers from clades A and C for clinical evaluation. *Vaccine* **39**, 3379–3387 (2021).
49. Sanders, R. W. et al. A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. *PLoS Pathog.* **9**, e1003618 (2013).
50. Kwon, Y. D. et al. A matrix of structure-based designs yields improved VRC01-class antibodies for HIV-1 therapy and prevention. *mAbs* **13**, 1946918 (2021).
51. Zhou, T. et al. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**, 811–817 (2010).
52. Charles, T. P. et al. The C3/465 glycan hole cluster in BG505 HIV-1 envelope is the major neutralizing target involved in preventing mucosal SHIV infection. *PLoS Pathog.* **17**, e1009257 (2021).
53. Bianchi, M. et al. Electron-microscopy-based epitope mapping defines specificities of polyclonal antibodies elicited during HIV-1 BG505 envelope trimer immunization. *Immunity* **49**, 288–300.e288 (2018).
54. Guenaga, J. et al. Well-ordered trimeric HIV-1 subtype B and C soluble spike mimetics generated by negative selection display native-like properties. *PLoS Pathog.* **11**, e1004570 (2015).
55. Wang, S. et al. HIV-1 neutralizing antibodies elicited in humans by a prefusion-stabilized envelope trimer form a reproducible class targeting fusion peptide. *Cell Rep.* **42**, 112755 (2023).
56. Li, W. et al. HIV-1 Env trimers asymmetrically engage CD4 receptors in membranes. *Nature* **623**, 1026–1033 (2023).
57. Zhou, T. et al. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445**, 732–737 (2007).
58. Kwong, P. D. et al. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678–682 (2002).
59. Zhou, T. et al. Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell* **161**, 1280–1292 (2015).
60. Irvine, E. B. & Reddy, S. T. Advancing antibody engineering through synthetic evolution and machine learning. *J. Immunol.* **212**, 235–243 (2024).
61. Xiao, Z. X., Miller, J. S. & Zheng, S. G. An updated advance of autoantibodies in autoimmune diseases. *Autoimmun. Rev.* **20**, 102743 (2021).
62. Wang, M., Patsenker, J., Li, H., Kluger, Y. & Kleinstein, S. H. Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic Acids Res.* **52**, 548–557 (2024).
63. Burbach, S. M. & Briney, B. Improving antibody language models with native pairing. Preprint at <https://arxiv.org/abs/2308.14300> (2023).

64. Bozhanova, N. G. et al. Computational identification of HCV neutralizing antibodies with a common HCDR3 disulfide bond motif in the antibody repertoires of infected individuals. *Nat. Commun.* **13**, 3178 (2022).
65. Schneider, C., Buchanan, A., Taddese, B. & Deane, C. M. DLAB: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* **38**, 377–383 (2022).
66. Hummer, A. M., Abanades, B. & Deane, C. M. Advances in computational structure-based antibody design. *Curr. Opin. Struct. Biol.* **74**, 102379 (2022).
67. Klein, F. et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* **153**, 126–138 (2013).
68. Bonsignori, M. et al. Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. *Immunol. Rev.* **275**, 145–160 (2017).
69. Karlsson Hedestam, G. B., Guenaga, J., Corcoran, M. & Wyatt, R. T. Evolution of B cell analysis and Env trimer redesign. *Immunol. Rev.* **275**, 183–202 (2017).
70. Perez, L.-H. et al. Direct bacterial killing in vitro by recombinant Nod2 is compromised by Crohn's disease-associated mutations. *PLoS ONE* **5**, e10915 (2010).
71. De Domenico, E. et al. Optimized workflow for single-cell transcriptomics on infectious diseases including COVID-19. *STAR Protoc.* **1**, 100233 (2020).
72. Dura, B. et al. scFTD-seq: freeze-thaw lysis based, portable approach toward highly distributed single-cell 3' mRNA profiling. *Nucleic Acids Res.* **47**, e16 (2019).
73. Perotti, M., Marcandalli, J., Demurtas, D., Sallusto, F. & Perez, L. Rationally designed human cytomegalovirus gB nanoparticle vaccine with improved immunogenicity. *PLoS Pathog.* **16**, e1009169 (2021).
74. Kschonsak, M. et al. Structural basis for HCMV Pentamer receptor recognition and antibody neutralization. *Sci. Adv.* **8**, eabm2536 (2022).
75. Kwon, Y. D. et al. Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env. *Nat. Struct. Mol. Biol.* **22**, 522–531 (2015).
76. Kong, R. et al. Antibody lineages with vaccine-induced antigen-binding hotspots develop broad HIV neutralization. *Cell* **178**, 567–584.e519 (2019).
77. Shu, Y. et al. Efficient protein boosting after plasmid DNA or recombinant adenovirus immunization with HIV-1 vaccine constructs. *Vaccine* **25**, 1398–1408 (2007).
78. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
79. Pancera, M. et al. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514**, 455–461 (2014).
80. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 486–501 (2010).
81. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. Sect. D. Struct. Biol.* **75**, 861–877 (2019).
82. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
83. Wu, T. T. & Kabat, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211–250 (1970).
84. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
85. Dunbar, J. et al. SAbPred: a structure-based antibody prediction server. *Nucleic Acids Res.* **44**, W474–W478 (2016).
86. Patil, I. Visualizations with statistical details: the 'ggstatsplot' approach. *J. Open Source Softw.* **6**, 3167 (2021).
87. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

Acknowledgements

The authors thank the study participants in Tanzania for donating blood samples for these studies. Sample collection was funded through this work is part of the IDEA project “Dissecting the Immunological Interplay between Poverty Related Diseases and Helminth infections: An African-European Research Initiative” (https://ec.europa.eu/research/health/infectious-diseases/neglected-diseases/projects/014_en.html) supported by the European Commission under the Health Cooperation Work Program of the 7th Framework Program (Grant 241642). The authors acknowledge the Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA for BG505 DS-SOSIP trimer; Alison Lin, David Vinyals Sales and Craig Fenwick from Lausanne University Hospital for advice and preliminary experiments with SOSIP trimer; Lausanne Genomic Technologies Facility, UNIL for next-generation sequencing of some of the samples and David Kalbermatter from the Dubochet Center for Imaging, Bern for cryo-EM grids data collection. This study was supported by the Swiss National Foundation (Grant number: 310030_20467) and intramural funding from Lausanne University—Lausanne University Hospital to L.P. Figure 1, created with BioRender.com, released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license.

Author contributions

M.F. and L.P. designed the project. P.N. with help from R.S. and R.R.W. performed and analyzed the experiments M.F., P.J., and R.G. computational work. S.O.D. and N.A.D.R. performed and analyzed the pseudo-viral neutralization assay experiments. D.D. set up of cryo-EM condition. M.M., O.L., C.D., Y.D.M., C.P., and M.P. samples or reagents. L.P. conceptualization, supervision, study design, data interpretation, and resources.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-49676-1>.

Correspondence and requests for materials should be addressed to Laurent Perez.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024