RAIN: a Machine Learning-based identification for HIV-1 bNAbs

Laurent Perez

Laurent.perez@chuv.ch

Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

Mathilde Foglierini

Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

Research Article

Keywords: HIV-1, broadly neutralizing antibody, Immune repertoire, Machine learning

Posted Date: March 8th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4023897/v1

Additional Declarations: The authors declare no competing interests.

Abstract

Broadly neutralizing antibodies (bNAbs) are promising candidates for the treatment and prevention of HIV-1 infection. Despite their critical importance, automatic detection of HIV-1 bNAbs from immune repertoire is still lacking. Here, we developed a straightforward computational method for **R**apid **A**utomatic **I**dentification of b**N**Abs (**RAIN**) based on Machine Learning methods. In contrast to other approaches using one-hot encoding amino acid sequences or structural alignment for prediction, RAIN uses a combination of selected sequence-based features for accurate prediction of HIV-1 bNAbs. We demonstrate the performance of our approach on non-biased, experimentally obtained sequenced BCR repertoires from HIV-1 immune donors. RAIN processing leads to the successful identification of novel HIV-1 bNAbs targeting the CD4-binding site of the envelope glycoprotein. In addition, we validate the identified bNAbs using *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.

Main text

More than 40 years after its identification, the human immunodeficiency virus-1 (HIV-1) remains a major global health concern[1]. The World Health Organization (WHO) estimates that in 2023 there were 38 million HIV-1 infected individuals worldwide, 1.5 million of 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 infection available[2]. Humoral immune response to HIV-1 targets the envelope (Env) protein of the virion, a trimeric membrane glycoprotein complex comprising gp120 and gp41[3]. However, the virus rapidly escapes immune control due to the exceptional Env glycoprotein diversity generated by HIV-1 error-prone replication machinery[4]. Moreover, additional mechanisms of immune evasion exist, such as heavy glycosylation of gp120, promoting a conformational masking of the receptor-binding site, enhancing immune evasion[5]. Screening of 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 (bNAbs), which can neutralize most viral strains at low concentrations *in vitro*[6]. Investigation of the development and structural properties of these bNAbs, revealed only a low level of sequence identity between them, but demonstrated that specific characteristics are associated with their function. For example, bNAbs have 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, bNAbs 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 treatment[9]. Passive transfer of bNAbs 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 human without notable adverse events or side effects[16,17]. BNAbs 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 gp120[18].

To date, the identification of bNAbs has required B cell isolation and clonal expansion from selected individuals possessing a serum 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 beneficiated from identified immune donors[19] and the development of high-throughput analyses of antibody repertoires by next-generation sequencing (**NGS**). Still, the number of identified HIV bNAbs remains relatively low, with only 250 of them reported[3,20]. Some bNAbs have been investigated in registered clinical trials, for prevention, as a component of long-acting antiretroviral therapy (ART), or as a component of intervention aimed at long-term drug-free remission of HIV[17,21,22]. Although, it is likely that the clinical success of bNAb passive immunization strategies will require a combination of antibodies to increase the overall breadth and potency against diverse HIV-1 isolates and to prevent the emergence of resistance[23]. The recent deployment of large datasets of human B cell repertoires on database repositories represents an opportunity for novel bNAbs identification assuming that computational tools for their automatic identification and classification are developed[24]. Artificial intelligence (AI)-based prediction tools to find the antibodies and antigens have been developed[25]. However, most of these tools rely on structural or amino acid sequence similarities of related antibodies to identify potential target proteins[26]. Nonetheless, despite important research and characterization efforts, a precise set of criteria required for classifying bNAbs versus non-bNAbs is still lacking.

Here, we developed a computational pipeline named RAIN for Rapid Automatic Identification of bNAbs from Immune Repertoire. 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 finally a R script converting the repertoire data into a features table for bNAbs prediction. We validated RAIN on previously identified bNAbs 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 cell from HIV-1 immune donors and performed single-cell BCR sequencing to demonstrate the method performance. Importantly, immune repertoire analysis of donors with a serum able to broadly neutralize different HIV-1 isolates led to the identification of three bNAbs, while none was detected in the repertoire of immune donors with sera that did not possess a broad neutralizing activity. The identified bNAbs were further characterized for their affinities to the envelope stabilized prefusion trimer BG505 DS-SOSIP, neutralizing activities and by cryoelectron microscopy (cryoEM) for one of them.

Results

Subset of discrete characteristics discriminate HIV-1 bNAbs from mAbs.

The automatic identification of HIV bNAbs 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 bNAbs isolated from chronically infected adults exhibit a signature of characteristic features, including high somatic hypermutations (SHMs), insertions or deletions (indels), long complementarity-determining region H3 (CDRH3), high potency, and broad viral neutralization breadth[3]. Moreover, the VRC01 class bNAbs, targeting the CD4bs, have also been shown to preferentially use specific germline alleles[27,28] and possess an unusually short CDRL3 of five amino acids, needed to contact gp120, while avoiding the glycan at position N267 in the D loop of gp120[29]. While bNAbs targeting the V1V2 apex use specific IGHV genes and together with bNAbs 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 bNAbs in a machine learning framework, could allow a rapid identification of bNAbs from an immune repertoire (**Figure 1)**. To identify predictors of HIV-1 bNAbs, we investigated specific features associated with these antibodies and inferred them from their highly diversified amino acid sequences. We collected and curated bNAbs sequences from the CATNAP (Compile, Analyze and Tally NAb Panels) database[32]. Data curation consisted of only considering human affinity matured sequences, removing incomplete or unpaired sequences obtained from CATNAP database (**Supplementary Table 1**). We obtained a total of 255 bNAbs paired sequences, binding the V1V2 apex (n=98), V3 glycan (n=56), CD4 binding site (n=54), gp120/gp41 interface (n=26) and MPER (n=21). Next, to create a dataset of paired BCR sequences that are unlikely to recognize an HIV antigen (hereafter named mAbs), we retrieved and curated paired antibodies sequences from ten healthy seronegative donors to obtain a total of 14'962 sequences (**Supplementary Table 2**). Following this step, we investigated if some of the bNAbs 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 **somatic hypermutation in the V gene** (**v**) or improbable acquired mutations in the framework regions only (**uv**), and the hydrophobicity of CDRH3[33,34] (**φ**) (**Figure 2a-e**). Interestingly, anti-CD4bs bNAbs analysis demonstrated a statistically higher **somatic hypermutation** frequency, a higher frequency of unconventional mutations (outside of the CDRs)[35], and a significantly shorter length of CDRL3 (**Figure 2a, b, e and Extended Data Fig 1a)** compared to the control mAbs reported in **Supplementary Table 2**. For the anti-MPER bNAbs, we observed a longer CDRH3, with higher hydrophobicity, and a higher mutation frequency in both V gene and framework (FRW) regions (**Figure 2a, b, c, d and Extended Data Fig 1b**). The bNAbs targeting the V1V2 apex showed a higher mutation frequency of V gene, but the difference was mainly due to a higher hydrophobicity of the CDRH3 and a longer CDRH3 (**Figure 2a-d and Extended Data Fig 1c**). BNAbs targeting the V3 glycan have higher frequency of mutations, and slightly higher hydrophobicity of the CDRH3 and a longer CDRH3 (**Figure 2a-d and Extended Data Fig 1d**). While bNAbs targeting the interface region, also demonstrated an increased frequency of mutations in the V gene and FWR regions (**Figure 2a, b and Extended Data 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 (**Figure 2f-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, both frequency of mutation in CDRs and framework regions were important, although 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 classify bNAbs from mAbs.

Algorithm selection and validation for the computational pipeline.

To further investigate the feasibility of an automatic identification of potential HIV-1 bNAbs, we decided to use different machine learning (ML) approaches. First, antibody sequences were converted to a list of values corresponding to the set of predictors identified previously. BNAbs sequences coming from CATNAP database were annotated using Igblast and the Immcantation 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 previously[39] and converted to a features table. For each antigenic site, bNAbs and mAbs were pooled as one dataset and subdivided into three: 60% as training set and 20% each as validation and testing set respectively. While the reported number of HIV-1 bNAbs is limited, a large quantity of mAb sequences is available. We thus decided to use first the anomaly detection algorithm (AD) for the automatic identification of 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 (**Extended Data Fig. 2a-e**), while evaluation of the AD performance, including computing of the area under the curve (AUC) was done with the test set (**Figure 3a, 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 (**Figure 3a**). 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 boundaries plot on the training dataset for bNAbs targeting the interface or the V1V2 apex (**Extended Data Fig 3a, b**). The receiver-operating characteristic (ROC) curve and corresponding AUC of 0.94 was obtained for V1V2 apex (**Extended Data Fig 3a**) and 0.9 for interface (**Extended Data Fig 3b**), indicating good classification performance for both antigenic sites. Furthermore, a measured AUC of 0.77 was obtained for bNAbs binding the CD4bs

(**Extended Data Fig 3c**). However, detection of bNAbs against other antigenic sites such as MPER (**Extended Data Fig 3c**), and V3 loop (**Extended Data 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 decision tree, the cost complexity pruning parameter (alpha) was set to zero (**Extended Data Fig 4**). Next, entropy was chosen as the quality measurement for the split in both DT and RF (further details are presented in the **Methods** section). Finally, we used the test datasets and evaluated performance metrics, including AUC, precision, recall and accuracy for the DT and RF models (**Figure 3a, c, and d**). The DT algorithm exhibited 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. Nevertheless, some expected characteristics were found, such as the frequency of mutation in V genes and unconventional mutations or the length of the CDR3 light chain, which has been described to be important for anti-CD4bs[40]. Similar parameters were observed for interface bNAbs, also characterized by their mutation frequency both conventional and unconventional. While the V1V2 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 weight (**Figure 3e**). The immunoglobulin variable VH5-51 gene segment was associated with bNAbs targeting the V3 glycan as previously reported for 35% of human anti-V3 bNAbs[41]. As a final validation step, we compared the prediction results of each algorithm. Altogether, we observed that the different methods (AD, DT, and RF) identified the same true positives, while there was minimal overlap in false positives (**Extended Data Fig 5**).

In an effort to combine the different ML algorithms used above, we chose to incorporate the Super Learner Ensembles algorithm (SL) as an additional validation step[42]. 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 (**Extended Data Fig 6a**) and achieved high performance for the MPER, V1V2 apex, and interface antigenic sites with a minimum AUC of 0.92 (**Extended Data Fig 6b**). In contrast, 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 (**Extended Data Fig 6a**). Based on the performance of our machine learning approach for Rapid Automatic Identification of bNAbs from Immune Repertoire (RAIN), we decided to use it on experimental samples in an effort to discovery new bNAbs.

Experimental validation of 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[43,44]. Interestingly, we observed that sera of donors 3, 11 and to some extent donor 9 had a broad neutralizing activity (**Figure 4a**). In contrast, sera from donors 1, 2, 5, 6, 7 and 8 were able to neutralize only one or two viruses (**Figure 4a**). Based on this result, we selected the donor 3 as test sample for bNAbs identification, while donors 1 and 2 were selected as negative control. 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 8,347 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 (**Extended Data Fig 7a**). To further confirm this result, we used the SL model, which identified thirteen potential bNAbs in donor 3: six predicted to bind to the CD4 binding site, one to V1V2 apex, and six interface binders (**Extended Data Fig 7b**). Interestingly, SL confirmed our predicted bNAbs, but also identified an anti-V1V2 apex binder in donor 2. Three potential bNAbs were constantly identified as CD4 binders (bNAb2101, bNAb4251, and bNAb1586) belonging to the VRC01-class and derived from the VH1-2*02 variable heavy chain gene segment (**Extended Data Fig 8**).

Binding and neutralization properties of the bNAbs.

To validate these findings, we cloned the potential bNAbs and some additional antibodies as negative control (hereafter referred as 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 trimers)[45,46], which is known to bind bNAbs that are representative of the majority of the known gp120 neutralizing antibody class[47,48]. Using biolayer interferometry (BLI), we detected a high-affinity interaction between all the identified bNAbs and SOSIP, characterized by an equilibrium dissociation constant (K_D) of 75nM, 3nM and 0.4nM for bNAbs 1586, 2101 and 4251 respectively. In contrast, no interaction could be detected between the control mAb and SOSIP (**Figure 4b**). To investigate the neutralization potency of our bNAbs, we sought of determining their IC_{50} using global HIV-1 panel strains on TZM-bl cells[43,44]. We observed a broad neutralization activity across tiers and viral clade for bNAb4251, with a geometric mean IC_{50} of 1.8µg/ml (**Figure 4c**). BNAb2101 could also neutralize different strains and specifically clade AE viruses, however its neutralization profile could not be considered as broad (**Figure 4c**). Finally, bNAb1586 was a relatively poor neutralizer, only able to inhibit the CNE55 strain at 38µg/ml (**Figure 4d**). 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, similarly to VRC01, which targets the CD4 binding site[49]. Since the potentials 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 BG505.W6M.C2 strain with residue T332N (C2) or N197, N276, N363, and N462 (gly4) and other mutations previously described[50] (**Figure 4d**). While the additional clade B viruses: JRCSF.JB (modified at D167N) to be more susceptible to VRC01 neutralization and YU2.DG, a tier 2 strain. Finally, clade C strains were also used as the glycan at 362 was naturally absent. The neutralization profile demonstrated a gain of potency specifically for the mutation surrounding CD4bs (**Figure 4d**).

Cryo-EM structures of BG505 SOSIP- FAb4251 complex.

To confirm the epitope and explore the binding mode of bNAb4251, we decided to perform cryo-electron microscopy (cryo-EM) of the antigen-binding fragment (Fab)4251 in complex with the soluble native-like trimer (BG505 DS-SOSIP)[51]. After several rounds of 2D and 3D classifications (**Extended Data Fig 9)**, we could segregate trimers with zero, or one Fab attached and solved the structure of the complex at the resolution of 3.7Å (**Figure 5a** and, **Supplementary Table 3**). As predicted by our method RAIN, Fab4251 interacts with the CD4bs of the trimer and makes multiple contacts with both heavy and light chain (**Figure 5a and b)**. In total, fifty-one residues of the Fab interact with fifty-six residues on gp120, to bury a surface area (bsa) of 950Å^2. The interaction is principally dictated by the heavy chain with 700Å^2 bsa, while the light chain buries 250Å^2 of the gp120 surface (**Figure 5c)**. The CDR-H2 makes most of the contact, totaling a bsa of 528Å^2, a binding mode that have similarity to the previously described interaction of the CD4 receptor with gp120 (**Figure 5d)**. 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[52-54] (**Figure 5d)**. Interestingly, H54 of CDR-H2 also mediates interaction with N368 and E370 of the "P43 cavity" located at the interface between the inner and outer gp120 domains (**Figure 5c** and **d).**

Previously reported bNAbs targeting the CD4-binding site (CD4bs) have been classified in 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, 8ANC131, VRC13 and VRC16)[55]. Structural investigation revealed that Fab4251 possesses an angle of approach similar to VRC01 (**Figure 5e),** a result in agreement with its CDR-H2 mediated contact on gp120, indicating that it belongs to the same antibody class. Moreover, hydrogen bounds are also present between the heavy chain R53, K62 and Q428, S460 of gp120, respectively (**Figure 5e)** and a salt bridge *between R71* and N368 of gp120. The CDR-H3 also contact the gp120 with N100 contributing a hydrogen bond with N279 in loop D of gp120, as it was already reported for other VRC01 class bNAbs **(Figure 5f).** 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-glycan[28], a feature also associated to 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 diagnostic and treatment for numerous immune-related or infectious diseases[56]. 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 arthritism, type 1 diabetes, multiple sclerosis, Grave's disease, Crohn's disease, and many others[57]. However, limitations exist and only 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[58,59]. Other efforts have focused on finding amino acid sequence similarity to an already known antibody. The similarity approaches led to important scientific and medical successed[60-62], but hold some limitations when 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 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 key factor to define a HIV-1 bNAbs. Former studies reported the presence of mutations in the frameworks of bNAbs and correlated with binding affinity to the CD4bs[34,63]. Our results suggest it is an important characteristic for all bNAbs. This can be interpreted as a consequence of the time needed for the maturation process 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[64,65]. The RAIN approach can achieve a precision of 1 for almost all antigenic sites and be applied easily on any immune repertoire or already isolated antibody sequences to identify HIV-1 bNAbs. Importantly, another unique 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 cryoEM. In summary, we believe that 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 mechanism of HIV-1 immune escape.

Material and Methods

Sample collection

Samples were obtained under study protocols approved by the Ethikkomission 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).

Serum IgG isolation

Serum samples from HIV-1-infected individuals were heat-inactivated at 56°C for 40 min and incubated with Protein G Sepharose (GE Life Sciences) overnight at 4 °C. IgGs were eluted from chromatography columns using 0.1Mglycine (pH= 3.0) into 0.1MTris (pH= 8.0). Buffer

was exchanged to PBS through Amicon 30 kDa spin membranes (Millipore). Concentrations of purified IgGs were determined by UV/Vis spectroscopy (A280) on a Nanodrop 2000 and samples were stored at -20 °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 20 min with the following fluorochrome-labeled mouse monoclonal antibodies: CD3-APC/Cy7 (dilution 1:40, clone HIT3a, catalogue no. 300317, BioLegend), CD27-Bv650 (dilution 1:50, clone O323, catalogue no. 302827, BioLegend), 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 CD3−CD20+CD27+IgG+ cells. FACS-sorted cells were collected in 6µl 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 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 System[66] (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 bar-coded 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 × 105 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)[67]. 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 (ThermoFisher 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 media. Heavy chain and light chain plasmids were complexed with Polyethyleneimine (ThermoFisher) and added to the cells. On day five, cells were cleared from cell culture media by centrifugation at 10,000*g* for 30 min and subsequently passed through a 0.45-μm filter. The supernatant containing the recombinant antibody was incubated with protein A resin (ThermoFisher) overnight at 4 °C. The resin was washed with 25 mL of phosphate-buffered saline (PBS). A total of 30 mL of 10 mM glycine pH 2.4, 150 mM NaCl were used to elute the antibody off the protein A resin. The acidic pH of the eluted antibody solution was increased to approximately 7 by the addition of 1M Tris pH 8.0. The antibody solution was buffer exchanged into PBS with successive rounds of centrifugation, filtered, 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, followed by size exclusion chromatography on a Superdex 16/600 HiLoad column (Cytiva).

Recombinant HIV-1 envelope SOSIP gp140 production

BG505 DS-SOSIP trimer[68] production and purification were performed as previously described[46]. 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 + F0HC 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 described[46]. Fractions containing theBG505 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[43,69,70]. Briefly, 293T-derived HIV-1 Env-pseudotyped virus stocks were generated by cotransfection of an Env expression plasmid and a pSG3ΔEnv backbone. Animal sera were heat-inactivated at 56°C for 1 hour and assessed at 8-point 4-fold dilutions starting at 1:20 dilutions. Monoclonal antibodies were tested at 8-point 5-fold dilutions starting at 50 µg/ml or 500 µg/ml. Virus stocks and antibodies (or sera) were mixed in a total volume of 50 µL and incubated at 37°C for 1 hr. TZM-bl cells (20 µl, 0.5 million/ml) were then added to the mixture and incubated at 37°C. Cells were fed with 130 µL 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 room temperature (RT) using an Octet K2 instrument (ForteBio). Protein A (Fortebio) biosensor probes were first equilibrated in reaction buffer for 60 s. IgGs were diluted to 10 µg/ml in reaction buffer and immobilized onto the protein A probes for 300 s, followed by a wash for 60 s in reaction buffer. The binding of SOSIP trimers to the IgGs was then measured at various concentrations for 300 s, followed by dissociation for 800 s in reaction buffer. Analysis was performed using the Octet software and GraphPad Prism version 9.0.

Cryo-EM sample preparation

BG505 DS-SOSIP trimers complexes were prepared using a stock solution of 2 mg/ml trimer incubated with a six-fold molar excess of bNAb4251. To prevent interaction of the trimer complexes with the air-water interface during vitrification, the samples were incubated in 0.085 mM n-dodecyl β-D-maltoside (DDM). Samples were applied to plasma cleaned QUANTIFOIL holey carbon grids (EMS, R2/2 Cu 300 mesh). The grid was blotted in an automatic plunge freezing apparatus Vitrobot MarkIV (Thermo Fisher, Hillsboro, USA) to control humidity and temperature.

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 4 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 approximately 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 including Motion correction and CTF determination[71]. Particle picking and extraction (extraction box size 350 $pixels^2$) were carried out using cryoSPARC[71]. 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 3.7 Å resolution (Fourier-shell coefficient (FSC) 0.143) with C1 symmetry. A model of a SOSIP trimer (PDB ID 4TVP)[72] or AlphaFold2 (ColabFold implementation) models of the 4251 Fab were fitted into the cryo-EM maps with UCSF Chimera. These docked models were extended and rebuilt manually with refinement, using Coot and Phenix[73,74]. Figures were prepared in UCSF Chimera, UCSF ChimeraX and Pymol[75]. The numbering of Fab 4251 is based on Kabat numbering of immunoglobulin models[76]. Buried surface area measurements were calculated within ChimeraX and PISA[77].

CATNAP sequences

For all antigenic sites, paired bNAb sequences were collected from the CATNAP database[32] as of 1st January 2022 as nucleotide and amino acid sequences. First, the 249 heavy chain and 240 light chain nucleotides sequences were annotated with Igblastn[36]. 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, bNAbs 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 1st January 2022). The 27 CATNAP antibodies with only the protein sequences available were annotated with IgBlastp followed by MakeDB.py from Change-O v1.2.0 (with the options igblast-aa –extended). In parallel, using the fasta protein sequences, ANARCI[78] was used to identify the junction region. As for nucleotide sequences, paired and annotated-CDR3 bNAbs were filtered in. In total, 255 bNAbs sequences were collected. Repartition of the antigenic site is as follows: 54 bNAbs target the CD4bs, 21 MPER, 98 V1V2, 56 V3, and 26 interface.

Paired B-cell receptor repertoires

For the training and evaluation of the machine learning models, paired BCRs repertoires of ten healthy donors were collected. The repertoires were obtained from various sources (**Extended Data Table 1**) and sequenced using 10X genomics technology. Annotation and processing of the sequences were done as previously described[39] 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 celescope 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 2’152 BCRs for D1, 6’195 BCRs for D2, 4’008 BCRs for B3, 3’794 BCRs for G3, 3’112 BCRs for S4, and 4’799 BCRs for G4.

Data pre-processing

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, aromatic residues having a 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 resides were set to zero. The values of all residues were summed up for each CDRH3. In addition, 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, dataset was partitioned into training, validation, and test sets with an 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 5 residues. For MPER, we included the combined feature VH1-69, VK3-20, and the GW motif in CDRH3. Additionally, CDRH3 hydrophobicity was added for MPER, V1V2, and V3. Lastly, CDRH3 length was incorporated for V1V2 and V3. Using the validation test, a multivariate normal random variable was calculated with the mutivariate_normal function from the scipy package v1.8.0 and used for setting the optimal Epsilon parameter ( ) minimizing the false positive numbers. The Epsilon value was set to 619.55 for CD4bs, 231501.41 for MPER, 866803.64 for V1V2, 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 pre-processing 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 (√122), and bootstrapping to build trees. Matplot 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 SuperLearner was created with the precision score as scorer parameter, a k-fold cross validation of 10 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 package[79]. The Complex Heatmap package was used for visualization[80]. 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.

Declarations

Data and 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 are available there. Raw sequencing data files for single-cell VDJ sequencing are available at GEO database: GSE229123. Cryo-EM map was deposited on EMDB: EMD-19665, with PDB accession number 8S2E.

Acknowledgements

We thank the study participants in Tanzania for their willingness to donate 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” (http://ec.europa.eu/research/health /infectious-diseases/neglected-diseases/projects/014_en.html) supported by the European Com

mission 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. This study was supported by the Swiss National Foundation (Grant number: 310030_20467) and intramural funding from Lausanne university-Lausanne University Hospital to LP. The authors acknowledge 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.

Author contributions

M.F. and L.P. designed the project. M.F., P.J., and R.G. computational work. P.N. with help from R.S. and R.R.W. performed and analyzed the experiments. S.O.D. and N.A.D.R. performed and analyzed the pseudoviral 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, resources.

Competing interests

The authors declare no competing interest.

References

1. Landovitz, R. J., Scott, H. & Deeks, S. G. Prevention, treatment and cure of HIV infection. *Nature reviews. Microbiology* **21**, 657-670 (2023). https://doi.org:10.1038/s41579-023-00914-1
2. Haynes, B. F. & Burton, D. R. Developing an HIV vaccine. *Science* **355**, 1129-1130 (2017). https://doi.org:10.1126/science.aan0662
3. Sok, D. & Burton, D. R. Recent progress in broadly neutralizing antibodies to HIV. *Nature immunology* **19**, 1179-1188 (2018). https://doi.org:10.1038/s41590-018-0235-7
4. Bailey, J., Blankson, J. N., Wind-Rotolo, M. & Siliciano, R. F. Mechanisms of HIV-1 escape from immune responses and antiretroviral drugs. *Curr Opin Immunol* **16**, 470-476 (2004). https://doi.org:10.1016/j.coi.2004.05.005
5. Malim, M. H. & Emerman, M. HIV-1 sequence variation: drift, shift, and attenuation. *Cell* **104**, 469-472 (2001). https://doi.org:10.1016/s0092-8674(01)00234-3
6. Liao, H. X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469-476 (2013). https://doi.org:10.1038/nature12053
7. Zhou, T. & Xu, K. Structural Features of Broadly Neutralizing Antibodies and Rational Design of Vaccine. *Advances in experimental medicine and biology* **1075**, 73-95 (2018). https://doi.org:10.1007/978-981-13-0484-2_4
8. Roskin, K. M. *et al.* Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nature immunology* **21**, 199-209 (2020). https://doi.org:10.1038/s41590-019-0581-0
9. Pantaleo, G., Correia, B., Fenwick, C., Joo, V. S. & Perez, L. Antibodies to combat viral infections: development strategies and progress. *Nature Reviews Drug Discovery* **21**, 676-696 (2022). https://doi.org:10.1038/s41573-022-00495-3
10. Shingai, M. *et al.* Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277-280 (2013). https://doi.org:10.1038/nature12746
11. 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). https://doi.org:10.1038/nature12744
12. Parsons, M. S. *et al.* Partial efficacy of a broadly neutralizing antibody against cell-associated SHIV infection. *Science translational medicine* **9** (2017). https://doi.org:10.1126/scitranslmed.aaf1483
13. Gautam, R. *et al.* A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges. *Nature* **533**, 105-109 (2016). https://doi.org:10.1038/nature17677
14. 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). https://doi.org:10.1016/j.cell.2014.07.043
15. Caskey, M., Klein, F. & Nussenzweig, M. C. Broadly neutralizing anti-HIV-1 monoclonal antibodies in the clinic. *Nat Med* **25**, 547-553 (2019). https://doi.org:10.1038/s41591-019-0412-8
16. Mendoza, P. *et al.* Combination therapy with anti-HIV-1 antibodies maintains viral suppression. *Nature* **561**, 479-484 (2018). https://doi.org:10.1038/s41586-018-0531-2

17. Gaebler, C. *et al.* Prolonged viral suppression with anti-HIV-1 antibody therapy. *Nature* **606**, 368-374 (2022). https://doi.org:10.1038/s41586-022-04597-1
18. McCoy, L. E. The expanding array of HIV broadly neutralizing antibodies. *Retrovirology* **15**, 70 (2018). https://doi.org:10.1186/s12977-018-0453-y
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). https://doi.org:10.1016/j.immuni.2019.02.008
20. Schriek, A. I., Aldon, Y. L. T., van Gils, M. J. & de Taeye, S. W. Next-generation bNAbs for HIV-1 cure strategies. *Antiviral research* **222**, 105788 (2023). https://doi.org:10.1016/j.antiviral.2023.105788
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). https://doi.org:10.1093/infdis/jiaa377
22. Sneller, M. C. *et al.* Combination anti-HIV antibodies provide sustained virological suppression. *Nature* **606**, 375-381 (2022). https://doi.org:10.1038/s41586-022-04797-9
23. Karuna, S. T. & Corey, L. Broadly Neutralizing Antibodies for HIV Prevention. *Annu Rev Med* **71**, 329-346 (2020). https://doi.org:10.1146/annurev-med-110118-045506
24. Marks, C. & Deane, C. M. How repertoire data are changing antibody science. *The Journal of biological chemistry* **295**, 9823-9837 (2020). https://doi.org:10.1074/jbc.REV120.010181
25. Kim, J., McFee, M., Fang, Q., Abdin, O. & Kim, P. M. Computational and artificial intelligence-based methods for antibody development. *Trends in pharmacological sciences* **44**, 175-189 (2023). https://doi.org:https://doi.org/10.1016/j.tips.2022.12.005
26. Akbar, R. *et al.* Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* **14**, 2008790 (2022). https://doi.org:10.1080/19420862.2021.2008790
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). https://doi.org:10.1126/science.1207227
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 U S A* **109**, E2083-2090 (2012). https://doi.org:10.1073/pnas.1208984109
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). https://doi.org:10.1126/science.aac5894
30. Liao, H. *et al.* Contribution of V(H) replacement products to the generation of anti-HIV antibodies. *Clinical immunology* **146**, 46-55 (2013). https://doi.org:10.1016/j.clim.2012.11.003
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). https://doi.org:10.1016/j.immuni.2022.09.001

32. Yoon, H. *et al.* CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research* **43**, W213-219 (2015). https://doi.org:10.1093/nar/gkv404
33. 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). https://doi.org:10.1016/j.chom.2020.01.027
34. Wiehe, K. *et al.* Functional Relevance of Improbable Antibody Mutations for HIV Broadly Neutralizing Antibody Development. *Cell host & microbe* **23**, 759-765.e756 (2018). https://doi.org:10.1016/j.chom.2018.04.018
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). https://doi.org:https://doi.org/10.1016/j.chom.2018.04.018
36. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research* **41**, W34-40 (2013). https://doi.org:10.1093/nar/gkt382
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). https://doi.org:10.1093/bioinformatics/bty235
38. Foglierini, M., Pappas, L., Lanzavecchia, A., Corti, D. & Perez, L. AncesTree: An interactive immunoglobulin lineage tree visualizer. *PLoS computational biology* **16**, e1007731 (2020). https://doi.org:10.1371/journal.pcbi.1007731
39. Phad, G. E. *et al.* Clonal structure, stability and dynamics of human memory B cells and circulating plasmablasts. *Nature immunology* **23**, 1076-1085 (2022). https://doi.org:10.1038/s41590-022-01230-1
40. 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). https://doi.org:https://doi.org/10.1016/j.immuni.2013.04.012
41. 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. *Molecular immunology* **46**, 917-926 (2009). https://doi.org:10.1016/j.molimm.2008.09.005
42. Laan, M. J. v. d., Polley, E. C. & Hubbard, A. E. Super Learner. *Statistical Applications in Genetics and Molecular Biology* **6** (2007). https://doi.org:doi:10.2202/1544-6115.1309
43. 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). https://doi.org:10.1128/JVI.02853-13
44. Schommers, P. *et al.* Restriction of HIV-1 Escape by a Highly Broad and Potent Neutralizing Antibody. *Cell* **180**, 471-489 e422 (2020). https://doi.org:10.1016/j.cell.2020.01.010
45. Chuang, G.-Y. *et al.* Structure-Based Design of a Soluble Prefusion-Closed HIV-1 Env Trimer with Reduced CD4 Affinity and Improved Immunogenicity. *Journal of Virology* **91**, 10.1128/jvi.02268-02216 (2017). https://doi.org:10.1128/jvi.02268-16

46. 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). https://doi.org:10.1016/j.vaccine.2021.04.063

47. 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). https://doi.org:10.1371/journal.ppat.1003618

48. 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). https://doi.org:10.1080/19420862.2021.1946918

49. Zhou, T. *et al.* Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**, 811-817 (2010). https://doi.org:10.1126/science.1192819

50. 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). https://doi.org:10.1371/journal.ppat.1009257

51. 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 reports* **42**, 112755 (2023). https://doi.org:10.1016/j.celrep.2023.112755

52. Li, W. *et al.* HIV-1 Env trimers asymmetrically engage CD4 receptors in membranes. *Nature* **623**, 1026-1033 (2023). https://doi.org:10.1038/s41586-023-06762-6

53. Zhou, T. *et al.* Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445**, 732-737 (2007). https://doi.org:10.1038/nature05580

54. Kwong, P. D. *et al.* HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678-682 (2002). https://doi.org:10.1038/nature01188

55. Zhou, T. *et al.* Structural Repertoire of HIV-1-Neutralizing Antibodies Targeting the CD4 Supersite in 14 Donors. *Cell* **161**, 1280-1292 (2015). https://doi.org:https://doi.org/10.1016/j.cell.2015.05.007

56. Irvine, E. B. & Reddy, S. T. Advancing Antibody Engineering through Synthetic Evolution and Machine Learning. *J Immunol* **212**, 235-243 (2024). https://doi.org:10.4049/jimmunol.2300492

57. Xiao, Z. X., Miller, J. S. & Zheng, S. G. An updated advance of autoantibodies in autoimmune diseases. *Autoimmunity Reviews* **20**, 102743 (2021). https://doi.org:https://doi.org/10.1016/j.autrev.2020.102743

58. 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 research* **52**, 548-557 (2024). https://doi.org:10.1093/nar/gkad1128

59. Burbach, S. M. & Briney, B. Improving antibody language models with native pairing. *arXiv preprint arXiv:2308.14300* (2023).

60. 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. *Nature communications* **13**, 3178 (2022). https://doi.org:10.1038/s41467-022-30865-9

61. 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). https://doi.org:10.1093/bioinformatics/btab660
62. Hummer, A. M., Abanades, B. & Deane, C. M. Advances in computational structure-based antibody design. *Current opinion in structural biology* **74**, 102379 (2022). https://doi.org:10.1016/j.sbi.2022.102379
63. 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). https://doi.org:10.1016/j.cell.2013.03.018
64. Bonsignori, M. *et al.* Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. *Immunological reviews* **275**, 145-160 (2017). https://doi.org:10.1111/imr.12509
65. Karlsson Hedestam, G. B., Guenaga, J., Corcoran, M. & Wyatt, R. T. Evolution of B cell analysis and Env trimer redesign. *Immunological reviews* **275**, 183-202 (2017). https://doi.org:10.1111/imr.12515
66. De Domenico, E. *et al.* Optimized workflow for single-cell transcriptomics on infectious diseases including COVID-19. *STAR Protoc* **1**, 100233 (2020). https://doi.org:10.1016/j.xpro.2020.100233
67. Dura, B. *et al.* scFTD-seq: freeze-thaw lysis based, portable approach toward highly distributed single-cell 3' mRNA profiling. *Nucleic acids research* **47**, e16 (2019). https://doi.org:10.1093/nar/gky1173
68. Kwon, Y. D. *et al.* Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env. *Nature structural & molecular biology* **22**, 522-531 (2015). https://doi.org:10.1038/nsmb.3051
69. Kong, R. *et al.* Antibody Lineages with Vaccine-Induced Antigen-Binding Hotspots Develop Broad HIV Neutralization. *Cell* **178**, 567-584.e519 (2019). https://doi.org:10.1016/j.cell.2019.06.030
70. Shu, Y. *et al.* Efficient protein boosting after plasmid DNA or recombinant adenovirus immunization with HIV-1 vaccine constructs. *Vaccine* **25**, 1398-1408 (2007). https://doi.org:10.1016/j.vaccine.2006.10.046
71. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature methods* **14**, 290-296 (2017). https://doi.org:10.1038/nmeth.4169
72. Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514**, 455-461 (2014). https://doi.org:10.1038/nature13808
73. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66**, 486-501 (2010). https://doi.org:10.1107/s0907444910007493
74. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta crystallographica. Section D, Structural biology* **75**, 861-877 (2019). https://doi.org:10.1107/s2059798319011471
75. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein science : a publication of the Protein Society* **30**, 70-82 (2021). https://doi.org:10.1002/pro.3943

76. 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. *Journal of Experimental Medicine* **132**, 211-250 (1970). https://doi.org:10.1084/jem.132.2.211

77. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology* **372**, 774-797 (2007). https://doi.org:10.1016/j.jmb.2007.05.022

78. Dunbar, J. *et al.* SAbPred: a structure-based antibody prediction server. *Nucleic acids research* **44**, W474-W478 (2016). https://doi.org:10.1093/nar/gkw361

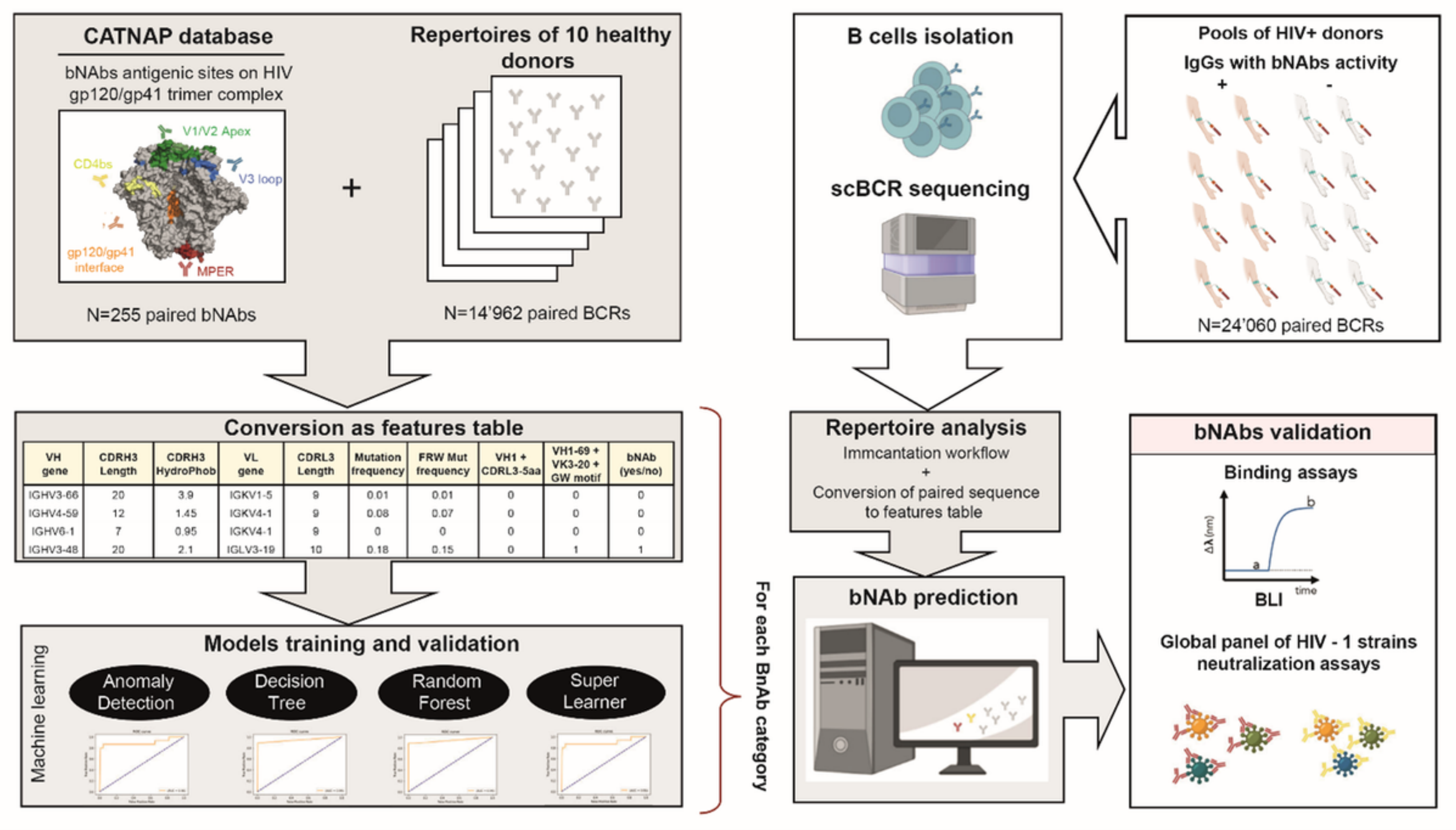
79. Patil, I. Visualizations with statistical details: The'ggstatsplot'approach. *Journal of Open Source Software* **6**, 3167 (2021).

80. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016). https://doi.org:10.1093/bioinformatics/btw313

Supplementary Table

Supplementary Table 1 is not available with this version

Figures



VH gene	CDRH3 Length	CDRH3 HydroPhob	VL gene	CDRL3 Length	Mutation frequency	FRW Mut frequency	VH1 + CDRL3-5aa	VH1-69 + VK3-20 + GW motif	bNAb (yes/no)
IGHV3-66	20	3.9	IGKV1-5	9	0.01	0.01	0	0	0
IGHV4-59	12	1.45	IGKV4-1	9	0.08	0.07	0	0	0
IGHV6-1	7	0.95	IGKV4-1	9	0	0	0	0	0
IGHV3-48	20	2.1	IGLV3-19	10	0.18	0.15	0	1	1

Figure 1

RAIN pipeline for automatic identification of bNAbs. Data collected from the CATNAP database (bNAbs) and healthy donor repertoires (mAbs) were converted as a feature 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 (illustrated by orange arm) or without (illustrated by white arm) sera broadly neutralizing activities. BCR sequences are processed by Immcantation workflow and analyzed as a features table. Next, the predicted bNAbs found by the four algorithms were produced and tested in neutralization and binding assays.

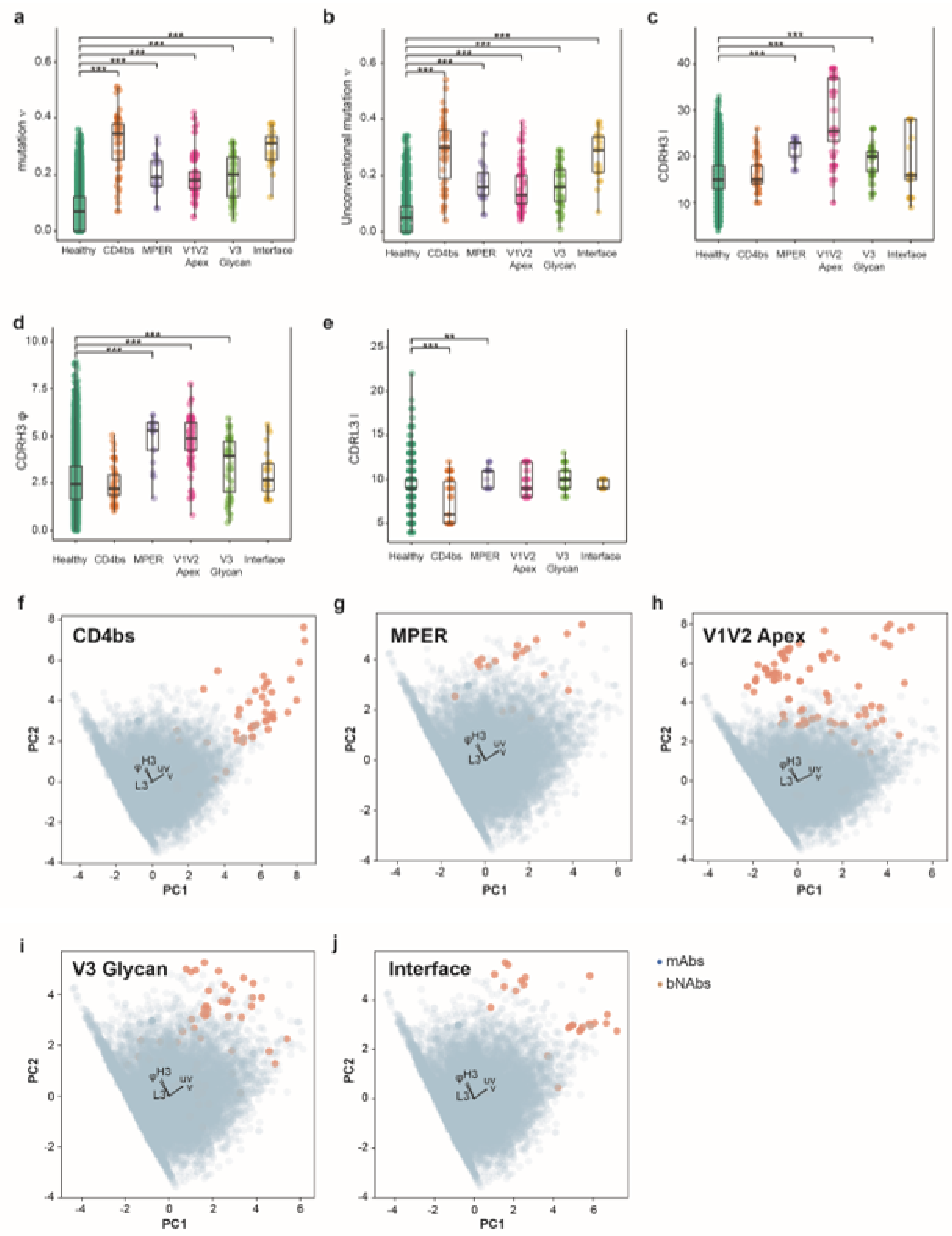
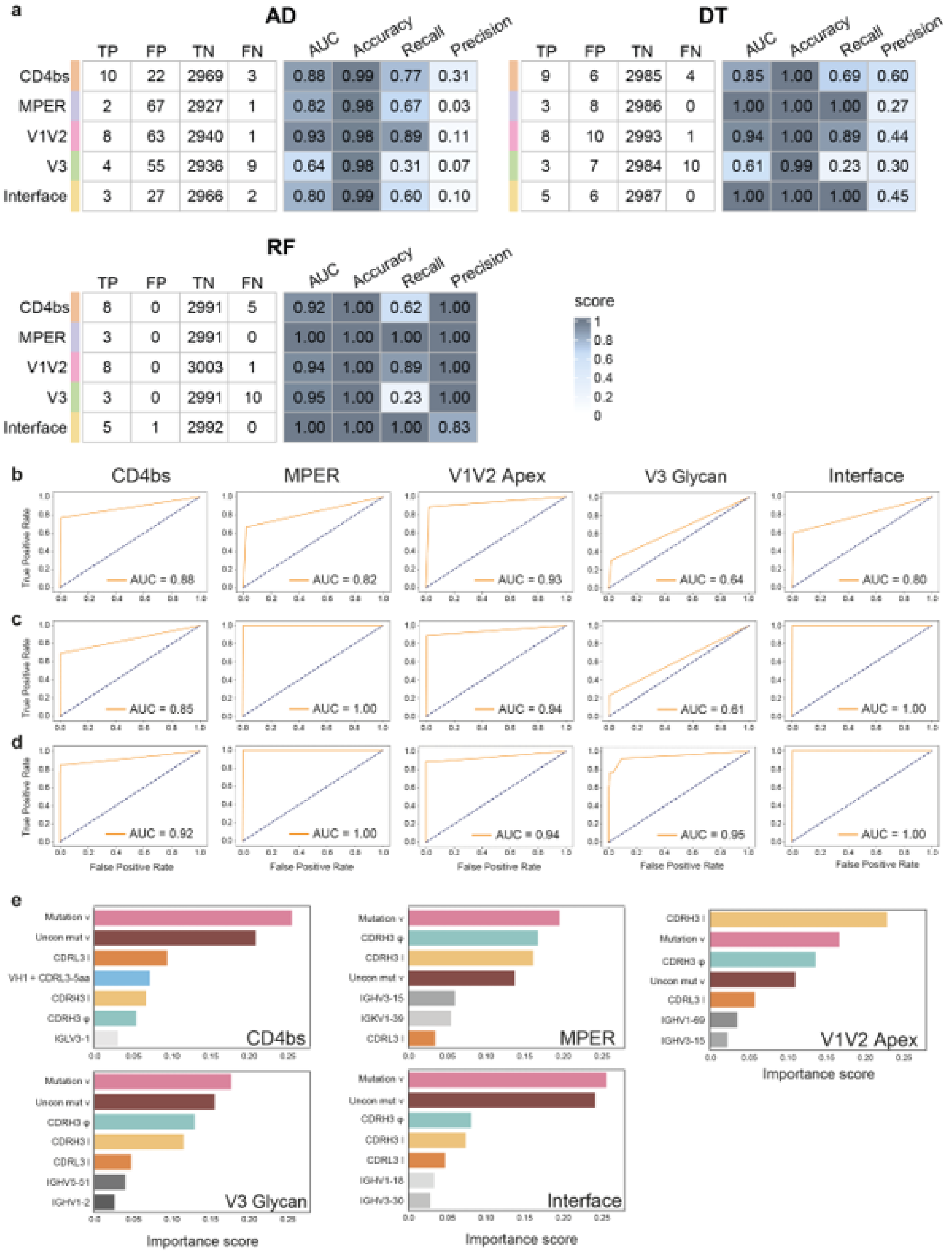


Figure 2

Characteristics discriminating HIV-1 bNAbs from mAbs. Specific properties of antibodies that allow differentiation between bNAbs and mAbs depending on the antigenic site. (**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 bNAbs category is represented by a single plot per antigenic site, (**f**)-CD4bs, (**g**)-MPER, (**h**) V1V2 apex, (**i**)-V3 glycan, and (**j**)-gp120/gp41 interface.



AD

	TP	FP	TN	FN	AUC	Accuracy	Recall	Precision
CD4bs	10	22	2969	3	0.88	0.99	0.77	0.31
MPER	2	67	2927	1	0.82	0.98	0.67	0.03
V1V2	8	63	2940	1	0.93	0.98	0.89	0.11
V3	4	55	2936	9	0.64	0.98	0.31	0.07
Interface	3	27	2966	2	0.80	0.99	0.60	0.10

DT

	TP	FP	TN	FN	AUC	Accuracy	Recall	Precision
CD4bs	9	6	2985	4	0.85	1.00	0.69	0.60
MPER	3	8	2986	0	1.00	1.00	1.00	0.27
V1V2	8	10	2993	1	0.94	1.00	0.89	0.44
V3	3	7	2984	10	0.61	0.99	0.23	0.30
Interface	5	6	2987	0	1.00	1.00	1.00	0.45

RF

	TP	FP	TN	FN	AUC	Accuracy	Recall	Precision
CD4bs	8	0	2991	5	0.92	1.00	0.62	1.00
MPER	3	0	2991	0	1.00	1.00	1.00	1.00
V1V2	8	0	3003	1	0.94	1.00	0.89	1.00
V3	3	0	2991	10	0.95	1.00	0.23	1.00
Interface	5	1	2992	0	1.00	1.00	1.00	0.83

Figure 3

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 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.

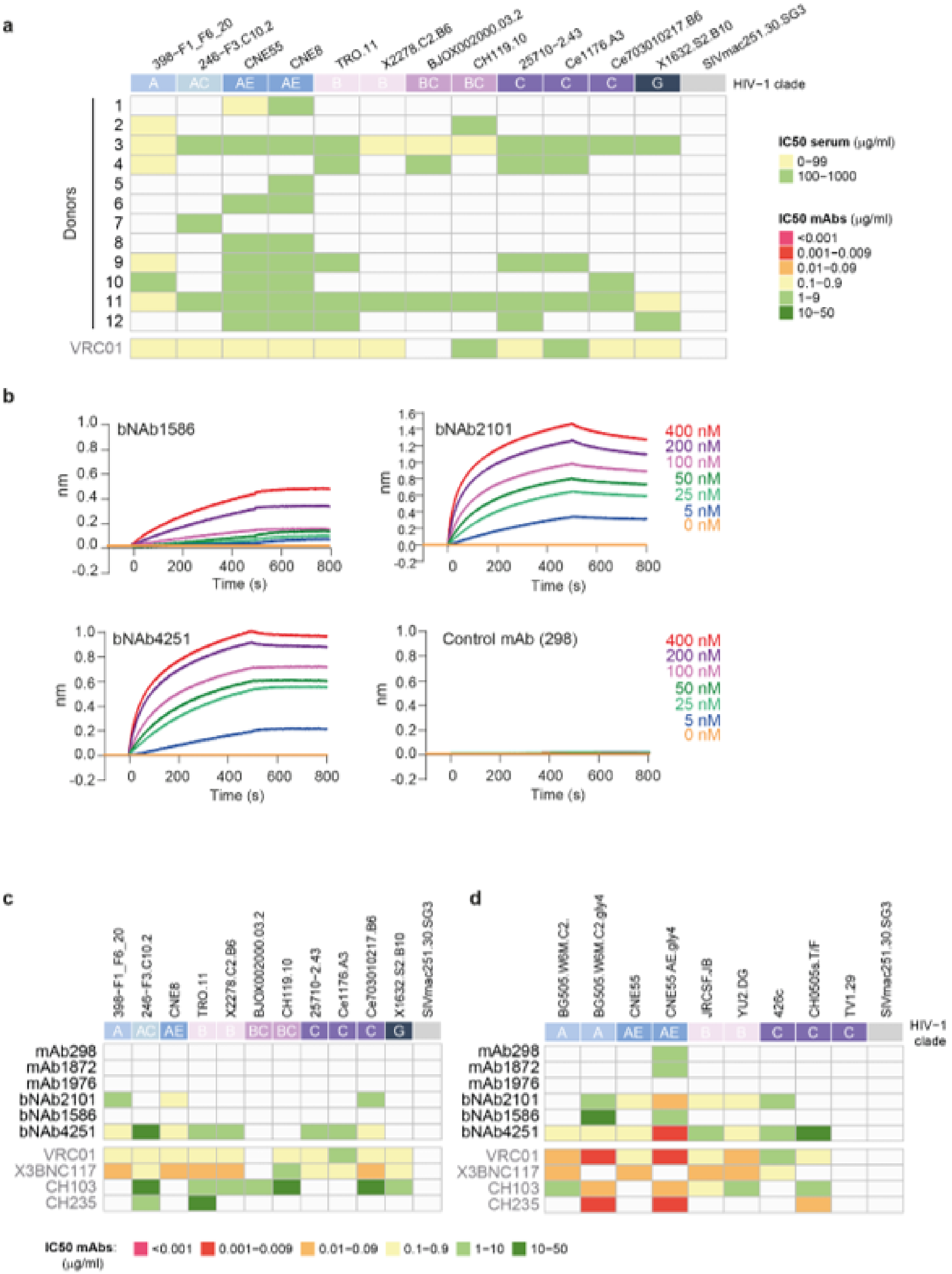


Figure 4

HIV Env binding and neutralization assays of serum and IgG samples. (**a**)-Neutralization assays were performed against twelve viruses from clades A, AC, AE, B, BC, C and G of tiers 2. The colors of the heatmap correspond to the IC_{50} of the sera in micrograms per ml. The SIVmac251.30.SG3 virus is used as negative control. (**b**)-Antibody SOSIP interactions were determined by biolayer interferometry (BLI). The mAbs or bNAbs were loaded on a protein G biosensor, dipped into 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. (**c**)-Neutralization assays were performed against twelve viruses from clades A, AC, AE, B, BC, C and G of tiers 2. The colors of the heatmap correspond to the IC_{50} in micrograms per ml, for each antibody. The SIVmac251.30.SG3 virus is used as negative. (**d**)-Neutralization assays were performed against glycan mutated viruses to support epitope mapping to the CD4 binding site.

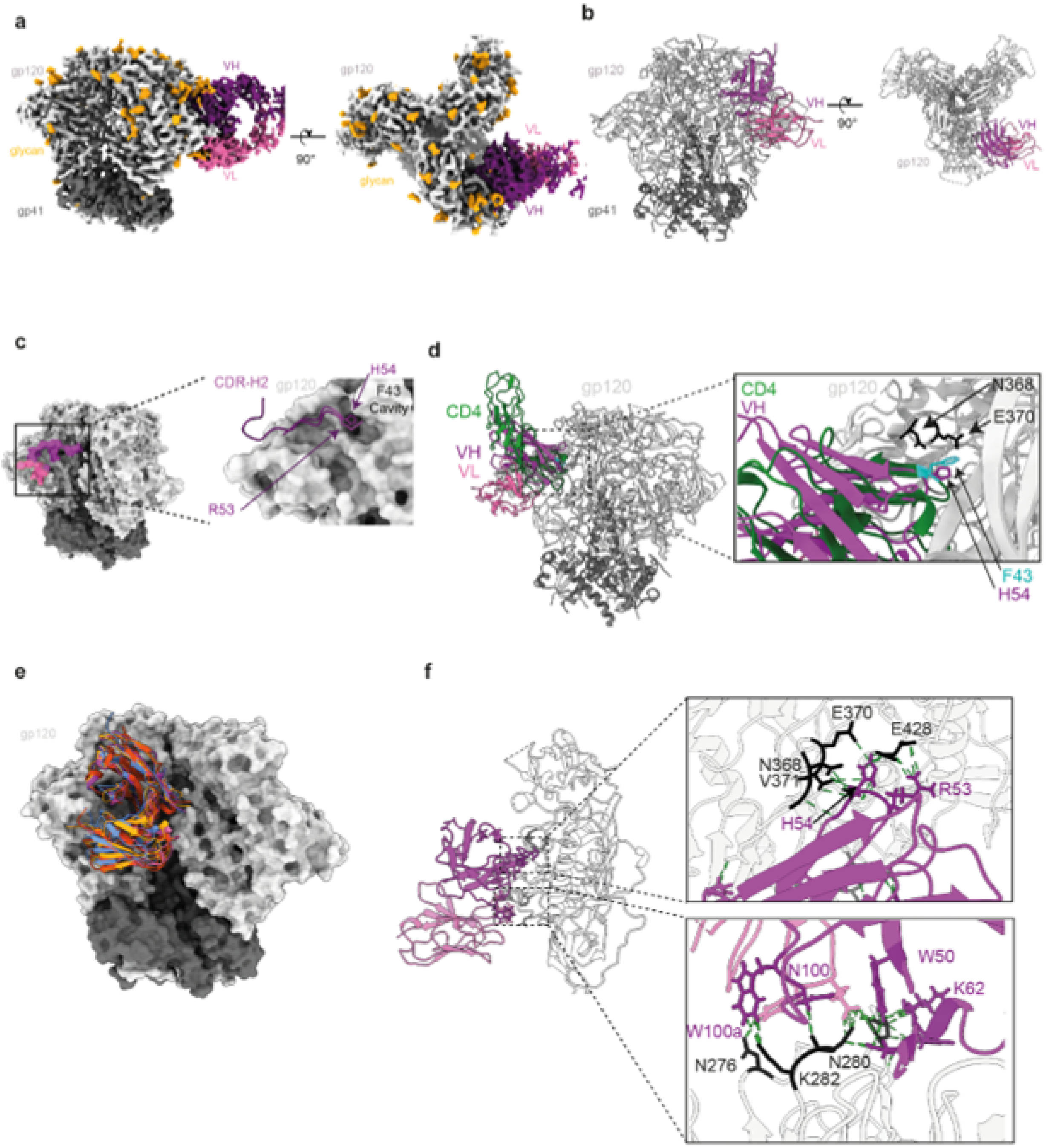


Figure 5

Cryo-EM map and structure of Fab SOSIP complexes. (**a**)-Side and top views of the cryo-EM density map of the Fab4251-DS-SOSIP complex, with gp120 in light grey, gp41 in dark gray, VH in violet and VL in pink. (**b**)-Atomic model of Fab4251-DS-SOSIP complex shown in cartoon representation. (**c**)-Foot print representation of the heavy and light chain binding surface on DS-SOSIP, colored according to a. Inlet on the right in represent the HCDR2 loop in violet, with H54 in the Phe-43 cavity. (**d**)-Overlay of CD4 receptor (green) bound to SOSIP (PDB.5U1F) and Fab4251 (violet). Inlet highlights positions N368, E370 on gp120 and F43 on CD4 and H54 of the VH. (**e**)-Overlay of VRC01 class antibodies on SOSIP with Fab4251 (violet), VRC01 (PDB.6V8X, green), PG04 (PDB.4I3S, red), and 3BNC60 (PDB.4GW4, orange). (**f**)-Contact residues at the Fab4251-gp120 interface. Contact residues are defined as two residues containing any atom within 4 Å of each other.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementoryInformation.docx