Design of immunogens to elicit broadly neutralizing antibodies against HIV targeting the CD4 binding site

Simone Conti[a], Kevin J. Kaczorowski[b,c], Ge Song[d,e,f], Katelyn Porter[d,e,f], Raiees Andrabi[d,e,f], Dennis R. Burton[d,e,f,g], Arup K. Chakraborty[b,c,g,h,i,1], and Martin Karplus[a,j,1]

[a]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; [b]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Scripps Consortium for HIV/AIDS Vaccine Development, The Scripps Research Institute, La Jolla, CA 92037; [e]IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA 92037; [f]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037; [g]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02139; [h]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; [i]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139; and [j]Laboratoire de Chimie Biophysique, Institut de Science et d'Ingénierie Supramoléculaires, Université de Strasbourg, 67000 Strasbourg, France

Contributed by Martin Karplus, January 7, 2021 (sent for review September 11, 2020; reviewed by Bruno E. Correia and Martin Weigt)

A vaccine which is effective against the HIV virus is considered to be the best solution to the ongoing global HIV/AIDS epidemic. In the past thirty years, numerous attempts to develop an effective vaccine have been made with little or no success, due, in large part, to the high mutability of the virus. More recent studies showed that a vaccine able to elicit broadly neutralizing antibodies (bnAbs), that is, antibodies that can neutralize a high fraction of global virus variants, has promise to protect against HIV. Such a vaccine has been proposed to involve at least three separate stages: First, activate the appropriate precursor B cells; second, shepherd affinity maturation along pathways toward bnAbs; and, third, polish the Ab response to bind with high affinity to diverse HIV envelopes (Env). This final stage may require immunization with a mixture of Envs. In this paper, we set up a framework based on theory and modeling to design optimal panels of antigens to use in such a mixture. The designed antigens are characterized experimentally and are shown to be stable and to be recognized by known HIV antibodies.

vaccine design | HIV | broadly neutralizing antibodies

Vaccines are the most important medical countermeasure for protecting entire populations against viruses, of which smallpox and measles vaccines are successful examples. In fact, a safe and effective HIV vaccine is considered to be the best way to end the global AIDS epidemic (1). However, how to produce a universal vaccine for highly antigenically variable viruses like HIV is a daunting and yet unsolved problem. The high variability of this virus allows it to elude the immune system, making the produced antibodies ineffective; that is, they are generally specific for a given strain of the virus but not for other strains resulting from mutation. In some cases, HIV-infected patients can elicit antibodies that can recognize and neutralize a broad range of different viral strains (2, 3). These broadly neutralizing antibodies (bnAbs) usually take a long time to appear naturally in infected patients and then only in a subset of such individuals.

The reason that bnAbs can arise is that even highly variable pathogens have regions with a well-defined, relatively conserved structure, which is required for their function. In HIV, entry depends on the trimeric spike exposed on the external lipid membrane of the virion, a heterotrimer formed by the gp120 and gp41 glycoproteins produced by posttranslational cleavage of a gp160 precursor. This protein binds to the CD4 coreceptor on CD4 T lymphocytes during HIV infection, and it has some relatively conserved regions that can be used as a target for bnAbs. Indeed, many bnAbs target the CD4 binding site (CD4bs) (4–8). If naive B cells that can bind to one of these relatively conserved regions can be expanded upon exposure to different variants of the virus, antibodies could evolve to better recognize the conserved portions, while avoiding the variable ones. The resulting antibodies can acquire breadth in this way, thereby becoming bnAbs. A successful vaccine would contain immunogens that can guide the immune system to produce bnAbs, rather than strain-specific antibodies.

In the past, numerous approaches for the development of an effective HIV vaccine have been tried. They include the use of cleverly chosen natural HIV proteins, the design of a consensus (9) or "center-of-tree" (10) antigens, and the creation of a mosaic protein from different HIV strains (11). All these methods used a single optimized antigen in the vaccine and were shown to be ineffective at eliciting bnAbs (12, 13). One possible reason for this is that, when exposed to a single antigen, the immune system will produce antibodies specific for that particular antigen, and neutralization escape variants can easily develop. A possible solution is to use more than one antigen in a vaccination protocol. This raises a number of questions: How many antigens are necessary? How different should they be from each other? And in what temporal order should they be administered? Answering such questions is far from trivial, in particular due to the limited mechanistic understanding of affinity maturation (AM) in vivo. Another problem is that bnAbs have an unusually high number of somatic mutations, not only in the complementarity-determining regions (CDRs) but also in the immunoglobulin

Significance

A solution to the global AIDS epidemic is to develop a vaccine against HIV. This has been difficult to achieve because the virus mutates rapidly, and the antibodies induced by vaccination would need to recognize many different HIV variants. We have developed a computational framework to design panels of antigens for eliciting broadly neutralizing antibodies (bnAbs) for an HIV vaccine. An important aspect of the method is its use of the gp160 fitness landscape, which measures the ability of the virus to tolerate mutations. Most designed antigens assembled as well-ordered native-like trimers with antigenic properties favorable for vaccine studies. These antigens can be used to make meaningful proposals for immunization schedules, a significant advance in HIV vaccine design.

Author contributions: R.A., D.R.B., A.K.C., and M.K. designed research; S.C., K.J.K., G.S., and K.P. performed research; S.C., K.J.K., G.S., K.P., R.A., D.R.B., A.K.C., and M.K. analyzed data; and S.C. and M.K. wrote the paper.

Reviewers: B.E.C., École Polytechnique Fédérale de Lausanne; and M.W., Sorbonne University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: arupc@mit.edu or marci@tammy.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2018338118/-/DCSupplemental.

Published February 26, 2021.

framework regions (7, 14). Recent computational data on the flexibility of the antibody and the need for framework mutations in the simulated AM showed how important it is for a vaccination protocol to have a specific antigen that can prime a good antibody precursor B cell receptor (BCR) (15). Moreover, it has been shown that putative precursors of known classes of bnAbs are generally not able to neutralize HIV or recognize envelope (Env), often due to clashes of the antibody with the glycosylation shield that protects the HIV Env protein (8, 16–18). For example, VRC01-class bnAbs are known to introduce a deletion or a mutation to a flexible glycine in the CDRL1 loop to avoid the glycan at N276 (19, 20).

The above discussion led to the proposal of a vaccination strategy consisting of three steps. First, a special purpose antigen is used to activate the correct naïve or precursor B cell (17, 21). Since this precursor will generally not bind to native HIV, as a second step, one or more antigens are used as intermediates to induce somatic mutations and to allow recognition of the native virus. In the third step, one or more antigens are used in a mixture or in sequence to increase the breadth of the antibody population (19, 22, 23). Implementations of the first and second steps have already been shown to be promising in experiments (16, 21, 24–27). However, much less is known about the third step. Some insights into this question can be obtained by in silico simulations of AM. Using coarse-grained models, it has been shown that, while administering a single mixture containing multiple antigens may induce too much frustration to lead to bnAbs formation, a sequential approach, in which antigens are administered one after another, seems to be more effective (23). It was also observed that the number of antigens required in a mixture is correlated with their sequence dissimilarity, and optimal breadth is obtained at an optimal number of antigens and dissimilarity (28). Given the coarse-grained nature of these studies, the actual antigen sequences to use in experiments cannot be obtained from them.

In this work, we focus on the third step of the proposed vaccination protocol. In particular, we derive a set of empirical rules and protocols to select an optimal panel of antigens to maximize the breadth of the produced antibodies upon AM. To be able to do so, it is essential to understand, at an atomistic level of detail, the role of each antigen amino acid in the antibody/antigen interaction. This aspect will be presented in the next section based on an analysis of the available crystallographic structures of bnAbs bound to the gp160 Env glycoprotein. However, the structures do not provide information concerning HIV stability and function. For example, generating antigen sequences by introducing purely random mutations will likely lead to sequences that are lethal for the virus and/or are not representative of HIV in vivo. To overcome this problem, it is useful to consider the structural data together with a model of the gp160 fitness landscape (29), which is a measure of the ability of HIV to tolerate mutations in its gp160 sequence to escape immune pressure. Structural and fitness information together provide a classification of the antibody/antigen interface and indicate the residues to mutate and the amino acids that are more probable at those positions.

While this analysis helps to reduce the number of antigen sequences to consider by highlighting the “hot spots” of antibody/antigen binding, it leaves open the question of how to select a combination of antigen sequences for use in a vaccine. Given rules of optimal sequence dissimilarity and optimal fitness according to the HIV landscape, a Pareto frontier approach will be described. It is able to select, from all possible panels of antigen sequences, the few that are predicted to best elicit antibodies with a broad activity spectrum. Experimental evidence of the viability of the designed antigens and of their immunogenic properties is presented in the final section.

Results and Discussion

Analysis of Experimental Structures of bnAbs Bound to the gp160 Env Glycoprotein. The RCSB Protein Data Bank (PDB) contains various experimentally determined structures of bnAbs bound to the gp160 Env glycoproteins. These can be sorted into three main categories according to the epitope on the gp160 to which antibodies bind: the CD4bs; the V1, V2, and V3 variable loops; and the membrane-proximal external region (MPER) (30) (Fig. 1). The first, the CD4bs, is the site gp120 uses to bind to the CD4 protein of CD4 T lymphocytes during viral infection. The residues and structure in this site are relatively well conserved, allowing the immune system to recognize them and to develop bnAbs that bind to this region. The second site, which contains the V1, V2, and V3 variable loops, is variable with regions of greater conservation, and antibody binding is affected by the heavy glycosylation in these loops. The third site is at the gp41/gp120 interface (MPER). Fewer bnAbs are known to bind to this region, and, as Fig. 1 suggests, the binding mode is not unique and is more diversified with respect to the CD4bs or the V1V2 or V3 binding. Among the three epitopes, the CD4bs follows most closely the methodology described in the Introduction, where antibodies can acquire breadth by recognizing functionally important gp120 residues.

Focusing on the CD4bs, 42 crystallographic structures of bnAbs bound to gp120 have been analyzed (see list in *SI Appendix*). Of these, 37 structures share essentially the same binding orientation with the VRC01 bnAb as example, while only five (corresponding to the antibodies b12, CH103, VRC13, VRC16, and HJ16) present major differences with the antibody rotated or shifted with respect to the reference VRC01 binding orientation. In this paper, we focus on the first group of 37 structures, as it is known that the VRC01-like family of bnAbs have breadth due to mimicking of CD4 binding (8). In the 37 crystallographic structures, we studied the binding orientation in more detail by determining which residues of the antibody and gp120 are in contact in the crystallographic structures and are thus used for binding; see *Methods*. This allows us to define a “usage” scale, ranging from 0% (never used) to 100% (used in all 37 structures), for each residue (Fig. 2 *A* and *B*). It is evident that the β15 and β23 beta sheets of the gp120 are used in binding for most analyzed antibodies, while part of the V5 and LD loops are only

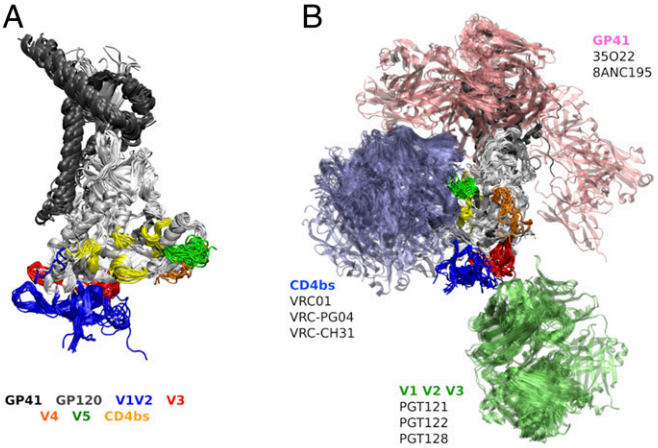

Fig. 1. (*A*) The 38 superimposed crystallographic structures of the gp120 (gray) and gp41 (black) complex. Different regions of the gp120 are highlighted in different colors: V1 and V2 variable loops are in blue, V3 is in red, V4 is in orange, V5 is in green, and the CD4bs is in yellow. (*B*) Superimposition of crystallographic structures of bnAbs bound to the gp120/gp41 complex. Blue, 24 structures bound to the CD4bs; green, 16 structures bound to the V3 glycan; red, 12 structures bound to the gp120/gp41 interface (MPER).

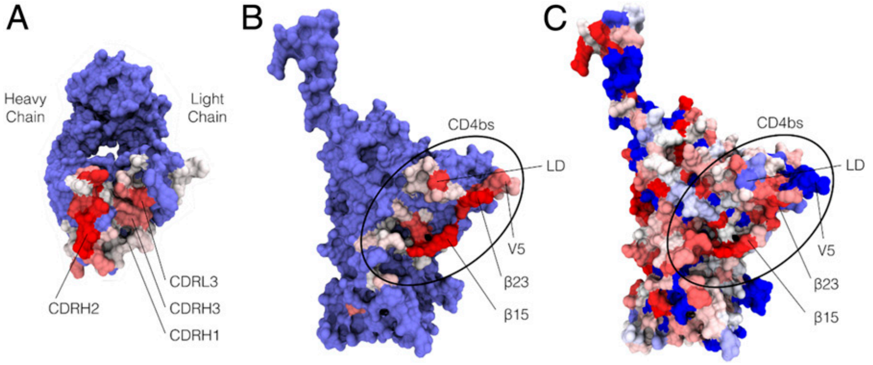

Fig. 2. Surface representations of the (*A*) VRC01 and (*B* and *C*) gp120 proteins. In *A* and *B*, the residues are colored according to the usage scale: residues in red are most frequently used in binding, those in blue are never used, and those in white are only seldom used. In *C*, the residues are colored according to the escape cost evaluated from the gp120 fitness landscape. Residues in red have high escape cost (approximately, highly conserved), while those in blue have very low escape cost (approximately, highly variable).

occasionally used. Looking at the antibody, the most prominent part used in binding is the CDRH2 loop, followed by CDRH3 and CDRL3. From the full complex shown in Fig. 3, it is clear that the CDRH2 loop of the antibody is "sandwiched" between the β15 and β23 beta sheets of gp120. Since these segments are the most used in all analyzed crystallographic structures, this is a key interaction for antibody/gp120 binding in the CD4bs, with the β15 and β23 beta sheets acting as anchor points for the antibody.

Comparison between the Usage Scale and the gp160 Fitness Landscape. A better understanding of the role of the residues in the CD4bs comes from a comparison between the usage of residues (defined above) and the fitness cost of evolving mutations at these residues. The latter can be obtained from the fitness landscape of the gp160 (29). In brief, the fitness landscape allows one to predict how difficult it is for HIV to escape immune pressure by evolving mutations in its gp160 sequence. The escape cost obtained from the fitness landscape accounts for epistatic couplings between mutations, and so describes the fitness penalty of evolving a mutation at a particular residue averaged over all possible sequence backgrounds (see *Methods*). The essential idea of our antigen design is to combine the residue usage map and the HIV fitness landscape to propose antigens that will elicit antibodies of high breadth from which viral escape will be minimized.

Fig. 2*C* shows gp120 residues colored according to their escape costs (red and blue indicate high and low escape costs, respectively). The residues in the β15 and β23 beta sheets have the highest escape costs. By contrast, the V5 residues, as well as those in the LD loop, are much easier to mutate. Comparing the escape costs to the usage scale (Table 1), the β15 and β23 beta sheets are extensively used in binding (usage between 79% and 93%), while the V5 and LD loops have the tendency to be avoided (usage between 14% and 74%). This supports the idea that antibodies evolved to recognize the high escape cost residues because of the high fitness penalties associated with accumulating mutations in those regions.

Based on these observations, we classify the residues in the CD4bs into three categories. Class 1 contains all residues that are highly used in binding to bnAbs and have a high escape cost ($E > 4$, where E is the fitness scale), such as the residues in the β15 and β23 beta sheets. This is the recognition site for the binding of the antibody. Variant antigen sequences that could be used as immunogens should not contain mutations in this class of residues. The eight residues in class 1 are numbers 365 to 368 (part of β15), 457 to 459 (β23), and 473. Residues in class 2 are often used but have a low escape cost ($E < 4$). These residues are usually spatially close to the class 1 residues and are always used in binding. This suggests that the common idea that there is a patch of conserved residues which is recognized by bnAbs is incomplete: The conserved residues are surrounded by and intermingled with highly mutable residues, and a bnAb has to "learn" to recognize the conserved ones and tolerate the variable ones. Variant antigens should bear different mutations at these residues to make sure the antibodies that develop upon immunization do not develop specificity for a particular amino acid. These residues are numbers 279 and 281 (LD), 371 (α3), and 460 to 463 (V5), for a total of seven residues. The residues in class 3 are seldom used and have a low escape cost. This class contains the majority of residues in the CD4bs. It is important to have them heavily mutated in the variant antigens to train the immune system to avoid this variable part and focus on the class 1 residues.

To select the residues to include in class 3, it is necessary to model which of these residues can come in contact with antibodies in the bound complex. We estimate this by carrying out all-atoms molecular dynamics simulations of the antibody/antigen complex. In contrast to the static structures resolved experimentally, with molecular dynamics, it is possible to simulate

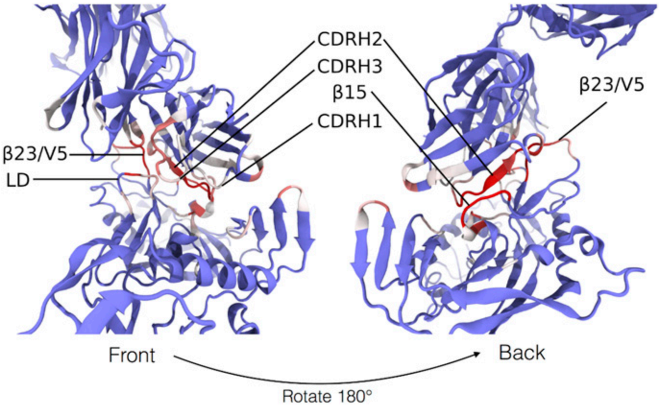

Fig. 3. Complex between the gp120 (at the bottom) and the VRC01 bnAb (at the top). The color code is the usage scale: residues in red are most frequently used in binding, those in blue are never used, and those in white are only seldom used. In the right side of the figure, it is evident how the CDRH2 loop of the antibody is "sandwiched" between the β15 and β23 loops of the gp120.

Table 1. Residues of BG505 SOSIP comprising classes 1, 2, and 3

Residue	Structural motif	Escape cost	In contact, %	VRC01, %	DRVIA7, %	VRC01GL, %	Amino acid
Class 1							
365	β15	4.18	81	100	85	98	90S
366	β15	6.44	93	100	100	100	100G
367	β15	6.84	93	100	100	100	100G
368	β15	7.35	88	100	100	100	100D
457	β23	6.57	79	100	100	100	99D
458	β23	6.69	86	100	100	100	100G
459	β23	4.95	93	100	100	100	96G
473		6.90	62	100	100	100	100G
Class 2							
279	LD	2.13	74	97	100	100	50D 47N
281	LD	3.69	36	100	100	100	79A 11T 6V
371	α3	3.12	62	100	100	100	89I 9V
460	V5	Nan	76	100	100	92	44N 11S 7E 7T 5D 5G
461	V5	Nan	60	100	100	100	28N 21T 11S 10D 10E 8G
462	V5	Nan	26	98	99	78	34N 29T 13S 7D 6E
463	V5	Nan	14	100	97	72	55N 12S 9T 6E
Class 3							
102	α1	2.76	0	76	99	58	84E 14D
105	α1	3.78	0	100	100	100	87H 12Q
194	V1V2	3.58	0	92	0	0	77I 8T 5R 5V
195	V1V2	4.07	0	50	0	0	65S 28N
236	β6	2.94	0	0	82	0	65T 22K 11S
275	LD	3.47	7	47	100	30	61E 14A 12D 5K
278	LD	2.39	7	99	100	100	70T 28S
283	LD	3.04	7	99	99	42	68T 15I 11N
353	LE	3.52	0	58	64	10	84F 10Y
360	β14	2.28	0	12	88	99	45V 27I 10A
364	β15	3.68	2	19	64	7	84S 10P
426	β20	3.23	0	1	91	0	67M 22R 9L
429	β21	3.54	21	51	97	73	76E 8K 6G
465	β24	Nan	0	100	100	92	60T 17N 9S
467	β24	1.99	0	100	96	93	53T 37I 10V
471	β24	3.73	10	76	32	6	80G 10A
474	α5	2.79	19	100	100	99	70D 29N
476	α5	2.86	0	100	100	96	75R 24K

The residue number follows the numbering in the 5D9Q crystallographic structure. The "structural motif" column indicates the secondary structure element of that residue, the "escape cost" is evaluated according to Eq. 4, the "in contact" is the fraction of the crystallographic structures of different bnAbs in which the residue is used in the binding, the "VRC01," "DRVIA7," and "VRC01GL" are the percentages of frames in the molecular dynamics simulation for which the residue is used in binding, and the last column is the population of each amino acid for the residue (e.g., "50D 47N" is read as "50% probability to be aspartic acid and 47% to be asparagine," where only amino acids with probability higher than 5% are reported).

the motion of the antibody/antigen complex and observe its flexibility. This makes it possible to determine which residues may come into contact only temporarily, providing a more comprehensive description of the binding site.

For these molecular dynamics simulations, the antigen BG505 SOSIP was chosen, for which crystallographic structures exist (19). SOSIP is used later as a template into which mutations are introduced to generate variant antigens that can be used as immunogens (see below). For the antibody, the VRC01 bnAb was chosen, as both sequence and crystallographic structures are available for the mature antibody (VRC01) (8), a putative germline (VRC01GL) (17), and an immature precursor (DRVIA7) (31). For each of the three antibody/antigen complexes, molecular dynamics simulations were carried out for 10 ns each; see *Methods*. The results of analyzing the generated ensemble of conformations are summarized in Table 1. Fifty-four residues of BG505 SOSIP were identified to be at the antigen/antibody interface. Fifteen of these residues were found previously to be members of class 1 or 2. Of the others, 21 have a high escape cost ($E > 4$), and so are unlikely to be mutated in an infecting strain; therefore, these will not be mutated in the variant antigens. The remaining 18 constitute the variable residues we defined in class 3, which the developing bnAbs need to learn to avoid, and so will contain mutations in the variant antigens.

Generating Variant Antigen Panels that Can Serve as Immunogens.

The next step in the design of immunogens is to create a panel of variant antigens with mutations in the 7 class 2 residues and 18 class 3 residues identified above. As a matter of practice, we consider panels composed of three antigens.

The first criterion we use to generate variant antigen sequences with mutations in the residues noted above is that they correspond to viable circulating strains. To ensure this, we consider only sequences publicly available within the Los Alamos National Laboratory (LANL) HIV Sequence Database (https://www.hiv.lanl.gov/). From the 20,043 sequences available, it is unfeasible to screen all possible combinations of three antigens ($1.3{\cdot}10^{12}$), and thus we limit the analysis to a sample of $\sim 1.13{\cdot}10^{11}$ randomly generated panels.

A second design consideration is the mutational distance between the antigens in the panel. We define the mutational distance between two antigens as the number of residues in class 2 or 3 at which the two sequences have different amino acids. It has been recently suggested that there is an optimal mutational distance between immunogens administered as a mixture that results in inducing antibodies with the maximum breadth (28): Too high a mutational distance leads to the extinction of germinal centers, while too low a distance results in induction of antibodies with low breadth. Although the coarse grain nature of that study did not allow a quantitative identification of a specific mutational distance, it is known experimentally that a mutational distance of 10 is too large (28). Thus, we elected to consider immunogen panels that have mutational distances with a mean of 5 ± 1 and variance of <1 among the sequences in the panel.

Considered together these two criteria, possible variant antigen panels are first created by selecting three antigens randomly from the set of 20,043 sequences in the LANL HIV Sequence Database. For each panel, the mean, μ, and variance, σ^2, of the pairwise mutational distances are calculated, and the panel is rejected if $|\mu{-}5| > 1$ or $\sigma^2 > 1$. Of the ~$1.13{\cdot}10^{11}$ random panels generated, $1.695{\cdot}10^8$ had the correct mutational distance. The number of panels generated by this method is too high to be practical, and additional criteria have to be added to select the most effective antigen panels.

The number of different residues is enforced to be around five, but this does not prevent substitution with chemically similar amino acids, for example, substitution of aspartic with glutamic acid, or of leucine with isoleucine. We wish to ensure that the mutations sampled within the class 2 and 3 residues are chemically dissimilar. To account for this, the average chemical similarity between each pair of sequences in a panel of variant antigens was estimated according to the pairwise similarity measure of McLachlan (32). This measure ranges from zero to six per residue, with large scores indicating a high degree of chemical similarity. The average McLachlan similarity across pairs of antigens was calculated for each immunogen panel and should be minimized in the choice of a variant antigen panel.

Given that even the best bnAbs are not able to neutralize every circulating HIV strain, we select antigens that are representative of natural variation, such that the vaccination can protect against strains most likely to be encountered in natural infection. For this, the HIV Env fitness landscape can be used, which makes the assumption, that seems to be validated by experimental tests, that the fitness of a strain is positively correlated with its prevalence within the global variation of HIV viral sequences (29). As described in *Methods*, a viral sequence with high fitness is assigned a low "energy." Thus, we seek to minimize the average energy of viral strains corresponding to the gp160 sequences of the panel of variant antigens.

Minimizing both the chemical similarity and the energy is a multiobjective optimization problem, for which the corresponding single-objective optima do not, in general, coincide. In other words, no panel exists that can simultaneously minimize both similarity and energy, and thus a trade-off has to be found. This type of problems can be addressed by constructing a "Pareto front" (33). In the context of our design problem, the Pareto front is the set of immunogenic panels for which no other panels exist that have both lower similarity and lower energy. For each of the generated random panels with an average distance around five, the average pairwise similarity and the average energy are calculated, and these values are input into a Pareto front calculation (see *Methods*). This process was repeated for the generated random sample of $1.695{\cdot}10^8$ panels. It resulted in a Pareto front containing only 21 panels (Fig. 4 and Table 2). The Pareto front converges rapidly as a function of sample size, as discussed in *Methods*.

A notable feature of the Pareto front is that the slopes near the ends change gradually; that is, for panels with the smallest energy, similarity decreases dramatically for slight increases in energy, while, for panels with the smallest similarity, slight decreases in similarity require large increases in panel energy. Panels at the center of the front are less sensitive. Moreover, the overall goal of the Pareto analysis is to find panels that minimize both the energy and the similarity at the same time. Looking at the central panels in Fig. 4, their energy and similarity are both below the energy and similarity of most of the random panels (the gray points). This cannot be said for the panels near the edges of the Pareto front. Thus, we focus on the panels near the middle of the front, since these come closest to simultaneously minimizing both panel similarity and energy. These are the best panels that can be generated and selected under the assumptions of this work.

Glycosylation. Glycosylation is an important aspect to consider when designing HIV antigens, because the presence of glycans effects the binding of antibodies. The gp160 Env glycoprotein is heavily glycosylated on its surface, with five sites in the vicinity of the CD4bs (N197, N234, N276, N363, and N462). Particularly problematic is the glycan on N276, which is accommodated by deletions in the CDRL1 loop in VRC01 class of bnAbs (19, 20).

Two strategies can be used to incorporate glycosylation in a sequential immunization protocol. In the first, the antigens are not glycosylated, and the degree of glycosylation increases with each successive immunization in the vaccination protocol. This would allow a greater variety of antibodies to bind to the antigens. They would be "filtered out" with the successive antigens as soon as they cannot accommodate the glycans. In the second strategy, the antigens are fully glycosylated from the beginning.

It is thus important to know whether the mutations selected in the 21 panels on the Pareto front affect in any way the glycosylation around the CD4bs. N-linked glycosylation needs a particular amino acid pattern: the target asparagine (N), which has to be part of a N-X-S/T pattern, where X can be any amino acid (apart from proline), and a serine (S) or threonine (T) must follow. Among the residues selected in our design protocol, there are four cases where we apply mutations that effect glycosylation patterns. First, we allow T236 in NGT to change, which eliminates the glycan at N234. Looking at the 21 panels on the Pareto front (Table 2), 12 panels (numbers 2 to 5, 7 to 11, 13, 14, and 20) have at least one sequence with a T236K or T236I mutation, which would prevent glycosylation at N234. Second, panel 8 has two sequences with the T278M mutation which breaks the NIT glycosylation pattern at N276. Third, one sequence in panel 9 has an N363P mutation which eliminates glycosylation at position N363. The last case of glycosylation is in the V5 loop, which is preserved in all panels on the Pareto front. Considering all cases,

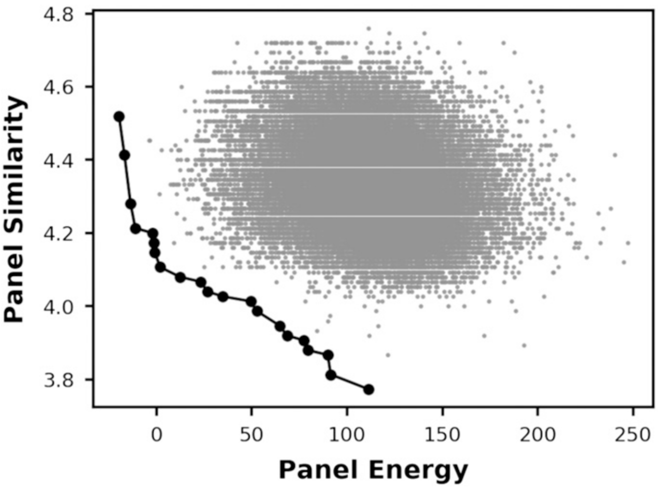

Fig. 4. Pareto front obtained by the analysis of $1.695 \cdot 10^8$ panels. The gray points are a sample of the panels generated, while the points in black are the panels on the Pareto front.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

IMMUNOLOGY AND INFLAMMATION

Table 2. Immunogenic panels obtained from the Pareto frontier analysis

Panel no.	LANL accession number	Sequence	Average energy	Average similarity
1	KJ704794	NAIEHISTDTTFVSMEIGDRNNTST	−19.63	4.52
	DQ410045	DAIEHISTQTIFVSMEIGDRNNENT		
	KR423589	NAIEQISTETVFVSMEIGDRNNTNT		
2	KR423584	NAIEQISIETVFVSMEIGDRNNTNT	−16.78	4.41
	KJ704794	NAIEHISTDTTFVSMEIGDRNNTST		
	DQ410045	DAIEHISTQTIFVSMEIGDRNNENT		
3	DQ410041	DAIEHISTQTIFVSMEIGDRNNENT	−13.58	4.28
	AY247219	DAIEHISKDTTFVSMEIGDRSNTST		
	KR423576	NAIEQISTETVFVSMEIGDRNNTNT		
4	AY247221	DAIEHISTDTTFVSMEIGDRNNEST	−11.10	4.21
	KR423573	NAIEQISTETVFVSMEIGDRNNTNT		
	AY247225	NAIEHISKETTFVSMEIGDRGTENT		
5	KJ704795	NAIEHISTDTTFVSMEIGDRNNKST	−2.18	4.20
	AY247225	NAIEHISKETTFVSMEIGDRGTENT		
	KR423589	NAIEQISTETVFVSMEIGDRNNTNT		
6	EU577245	NAIEQINTETTFISMEIGDRNPNGT	−1.21	4.17
	KJ704795	NAIEHISTDTTFVSMEIGDRNNKST		
	KR423304	NAIEQISTETVFVSMEIGDRNNTNT		
7	KR423577	NAIEQISIETVFVSMEIGDRNNTNT	−0.88	4.15
	AY247225	NAIEHISKETTFVSMEIGDRGTENT		
	KJ704795	NAIEHISTDTTFVSMEIGDRNNKST		
8	DQ410105	DAIEHISKAMIFVSMEIGDRTNEST	1.93	4.11
	DQ410045	DAIEHISTQTIFVSMEIGDRNNENT		
	AY247218	DAIEHISTDMTFVSMEIGDRNKSET		
9	KJ704792	DAIEHINKDTTFVSMEIGDRNXEXT	12.36	4.08
	DQ410043	DAIEHISTQTIFVSMEIGDRNNENT		
	KF384809	DAIEHISKATIFVPMEIGDRKNEST		
10	KR423574	NAIEQISTETVFVSMEIGDRNNTNT	23.30	4.07
	AY247225	NAIEHISKETTFVSMEIGDRGTENT		
	HQ217532	NAIEHISTKTTFSSMEIGDRTNGNT		
11	AY247225	NAIEHISKETTFVSMEIGDRGTENT	26.75	4.04
	KR423280	NAIEQISTETVEVSMEIGDRNNTNT		
	HQ217528	NAIEHISTETTFSSMEIGDRTNGNT		
12	KR423280	NAIEQISTETVEVSMEIGDRNNTNT	34.70	4.03
	HQ217523	NAIEQISTETTFSSMEIGDRTNGNT		
	EU577271	NAIEQINTETTFISMEIGDRNPNGT		
13	GU728339	NAIEHISKETTFVSREIGDR-NTNT	49.31	4.01
	EU576921	NAIEHINTETTFVSMEIGDR-KNST		
	KR423573	NAIEQISTETVFVSMEIGDRNNTNT		
14	GU728339	NAIEHISKETTFVSREIGDR-NTNT	52.56	3.99
	EU576898	NAIEHINTETTFVSMEIGDR-NNST		
	KR423280	NAIEQISTETVEVSMEIGDRNNTNT		
15	EU576922	NAIEHINTETTFVSMEIGDR-NNST	64.60	3.95
	AB588218	NAIEHINTETTFVSRETADR-GGST		
	EU578613	NAIEQINTETTFSSMETGDR-PNDT		
16	AY357345	NAIEHISTETTFVSMETGDR-TNNN	68.41	3.92
	EU744103	NAIKHRSTETIFVSMEIGDR-DNGT		
	EU576921	NAIEHINTETTFVSMEIGDR-KNST		
17	AY357517	NAIEHISTETTFVSMEIGDR--NTE	77.12	3.91
	EU576940	NAIEHINTETTFASMETGDR-VNGR		
	EU578617	NAIEQINTETTFSSMETGDR-PNDT		
18	AF025757	NAIEHINTETTEVSREIGDR--NTE	79.24	3.88
	EU576921	NAIEHINTETTFVSMEIGDR-KNST		
	EU576946	NAIEHINTETTFASMETGDR-VNGR		
19	AF025757	NAIEHINTETTEVSREIGDR--NTE	89.85	3.87
	KC473825	NAIEHISTETIFISMTIGDR--NSE		
	EU576910	NAIEHINTETTFVSMEIGDR-NNST		
20	GU562101	NAIEHISKETTFVSMEIGDR-ENNT	91.35	3.81
	AF277064	NAIDHINTETTFVSMEIGDR--TNT		
	AF025757	NAIEHINTETTEVSREIGDR--NTE		
21	AF277065	NAIEHINTETTFASMETGDR-VNGR	111.07	3.77
	AF025757	NAIEHINTETTEVSREIGDR--NTE		
	FJ469716	NAIEHIHTETTFVSMETGDR--NEK		

Amino acids in sequence are in order 279 281 371 102 105 194 195 236 275 278 283 353 360 364 426 429 467 471 474 476 460 461 462 463 465.

nine panels do not affect glycosylation; these are panels 1, 6, 12, 15 to 19, and 21.

Here, we assume that the chosen antigens need to be as similar as possible to the native virus, and mutations that disrupt glycosylation patterns have to be avoided. Panel 12 was selected, therefore, for experimental evaluation.

Production and Characterization of Antigenic Constructs. In order to produce antigens that can be used in a potential vaccine regimen, the mutations identified in the optimal panel 12 were introduced into the wild-type BG505 SOSIP.664 Env trimer (BG-WT), which is stable in solution and is immunogenic in animal models (34–37). Panel 12 is composed of antigens derived from GenBank sequences corresponding to the following accession numbers: EU577271, HQ217523, and KR423280 (Table 3). The critical CD4bs residues of these three antigen sequences were incorporated onto the BG-WT SOSIP background sequence, resulting in three new antigen sequences designated as SOSIPs of BG-EU, BG-HQ, and BG-KR (Fig. 5). These modified CD4bs trimer variants were expressed by cotransfection of plasmids encoding the soluble SOSIP trimers and furin into mammalian HEK293F cells. The soluble antigens were then affinity purified and separated by size exclusion chromatography (SEC). As expected, the SEC profiles revealed that BG-WT predominantly exists in trimer form. Two of the trimer variants, BG-EU and BG-HQ, exhibited substantial proportions of the protein that assemble as well-folded trimers, and can thus be recovered as soluble trimers in testable amounts. Only a small proportion of trimer population was observed for the BG-KR variant (*SI Appendix*, Fig. S1).

To determine the antigenic properties, the trimers were tested with a range of HIV Env-specific mAbs targeting various epitopes, by biolayer interferometry (BLI) and compared with the BG-WT trimer. BLI revealed that the levels of binding responses of various mAbs with the trimer variants varied compared to BG-WT trimer, but the overall binding kinetics were similar, except for some of the CD4bs mAbs (Fig. 5). The CD4bs mAbs 3BNC117, PGV04, HJ16, b12, b6, and F105 showed differing binding kinetics for the trimer variants compared to BG-WT trimer, suggesting that the substitutions at the selected CD4bs residue positions alter their interactions with these antibodies. The overall antigenic profiles looked favorable for immunization, and these profiles are expected to further improve for PGT145 Ab affinity purified material, as this quaternary trimer-specific bnAb enriches for well-ordered homogenous trimers populations (38).

As described in the Introduction, a successful HIV vaccine is likely to involve multiple stages, starting with a priming agent to activate a desired antibody precursor, followed by a set of intermediate antigens to shepherd AM appropriately, terminating with a set of native-like antigens to increase the breadth of the immune response. Although we focus our design on the final immunization step, the identified mutations were also applied to antigen designs for the second stage of affinity improvement; that is, the mutations corresponding to the EU, HQ, and KR antigens were introduced into the core and core-GT3 constructs; sequences are reported in *SI Appendix*. These are antigenic constructs from the literature, which are known to recognize the antibody germlines elicited by the priming agents (24).

The core and core-GT3 versions of the proteins were expressed by transfection of the plasmids encoding these constructs into HEK293F cells. The proteins were then purified from the culture supernatants and separated by SEC. The analysis of the core and the core-GT3 forms of the variants revealed consistent biochemical properties comparable to those of the corresponding BG-WT forms. As done for the trimers, the core and core-GT3 variants were tested with a range of HIV Env CD4bs-specific mAbs by BLI, in order to determine their antigenic properties (Fig. 6). BLI revealed substantially varied binding profiles of CD4bs-specific mAbs with WT compared to the core and core-GT3 antigen variants. As expected, the core-GT3 variants showed much stronger binding with the CD4bs mAbs as compared with the core versions, as the core-GT3 antigen versions contain additional mutations that enhance binding by CD4bs bnAbs (24). The binding kinetics of the VRC01 bnAb were comparable to those for the WT-core-GT3 and its corresponding variants. However, the bnAbs 3BNC60 and PGV04 exhibited reduced binding to the core-GT3 variants, especially with the BG-EU, suggesting differential epitope recognition by these bnAbs. Overall, while a range of activities was observed, all core and core-GT3 versions exhibited favorable biochemical properties.

The selected mutations in panel 12 were introduced into both the trimeric and intermediate monomeric antigenic constructs, and most of the antigens were successfully expressed. Their ability to bind to known antibodies was tested, and all showed favorable properties.

Conclusions

Finding an HIV vaccine is considered to be an essential step toward eradicating HIV. Recent efforts are pursuing a three-stage vaccine: First, precursors are primed using a specifically designed construct, then intermediate immunogens are used to adapt the antibodies to the peculiarities of native HIV, and, finally, a mixture or sequence of native-like antigens is used to improve the breadth of the immune response.

Here, we combine atomistic descriptions of HIV antibody/antigen binding with the fitness landscape of gp160 Env glycoprotein to optimize antigen panels for use as immunogens in a vaccine. First, we classify the residues at the antibody/antigen interface based on their involvement in the binding and their contribution to viral fitness. Three classes of residues are defined: 1) those that are important for the binding and are conserved, 2) those that are important for the binding and are variable, and 3) those that seldom interact with the antibodies and are variable. Residues in the first class form the recognition site, which is constant for any antigen. Residues in the second and third classes are the reason it is very difficult to produce antibodies with a high breadth: These residues are extremely mutable, and antibodies need to adapt to and tolerate their variability to acquire breadth. Based on this analysis, we developed a protocol to design panels composed of three antigens such that the chemical similarity between the three is minimized, and their fitness is maximized. This produces panels of antigens

Table 3. Selected immunogenic panel obtained from the Pareto frontier analysis

LANL accession number	Sequence	Fitness energy	Similarity
BG505 SOSIP	EHINTETNANFRSVMRSTNSTTGDR		
KR423280	E**Q**I**S**TETNA**VEV**S**I**M**ENNTN**T**I**GDR	34.70	4.03
HQ217523	E**Q**I**S**TETNA**T**F**S**S**I**M**ETNGN**T**I**GDR		
EU577271	E**Q**INTETNA**T**F**I**S**I**M**ENP**N**G**T**I**GDR		

Amino acids in sequence are in order 102 105 194 195 236 275 278 279 281 283 353 360 364 371 426 429 460 461 462 463 465 467 471 474 476. Residues that differ from BG505 reference sequence are boldfaced.

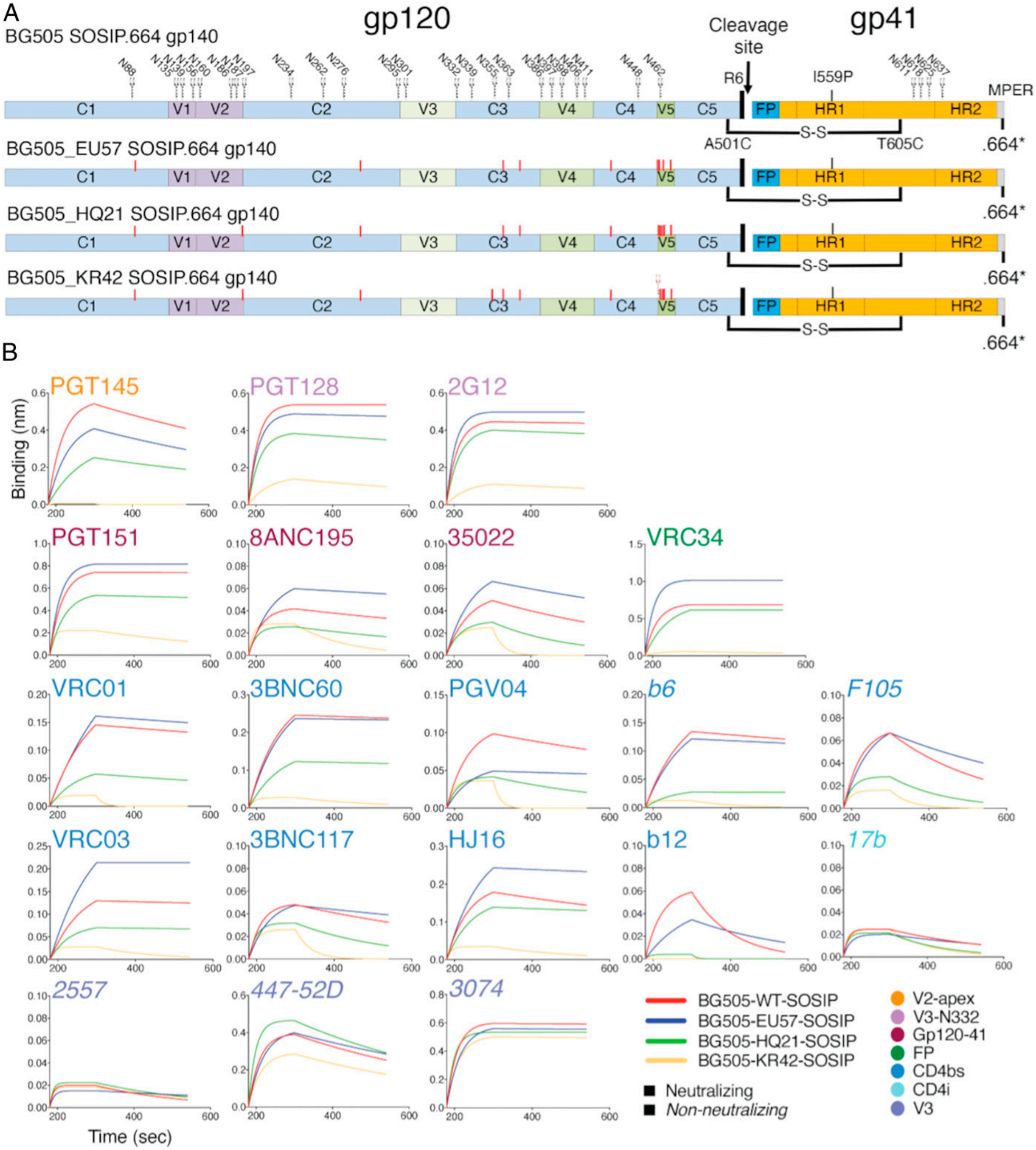

Fig. 5. (*A*) Schematic showing the design of BG505 SOSIP.664 gp140 soluble trimer and its CD4bs SOSIP trimer variants (BG-EU, BG-HQ, and BG-KR). The gp120 and the gp41 regions and the soluble SOSIP.664 trimer-stabilizing modifications are indicated. The potential N-linked glycan site positions for each NXT or NXS residue are labeled according to the HIV HXB2 numbering scheme. The substitutions in the CD4bs SOSIP variants at various gp120 positions corresponding to BG505 Env are depicted as vertical red lines. (*B*) BLI binding of HIV Env-specific mAbs to BG505 SOSIP trimer and its CD4bs trimer variants. HIV Env mAbs targeting various epitope specificities (V2 apex, V3-N332, gp120-41 interface, FP, CD4bs, CD4i, and linear V3) were tested for binding. BLI binding curves (association, 120 s [180 to 300]; dissociation, 240 s [300 to 540]) of HIV Env mAbs to BG505 SOSIP trimer and its CD4bs trimer variants. The mAbs were immobilized on human IgG Fc capture biosensors and the trimer proteins were used as analytes. The binding response is shown in nanometers.

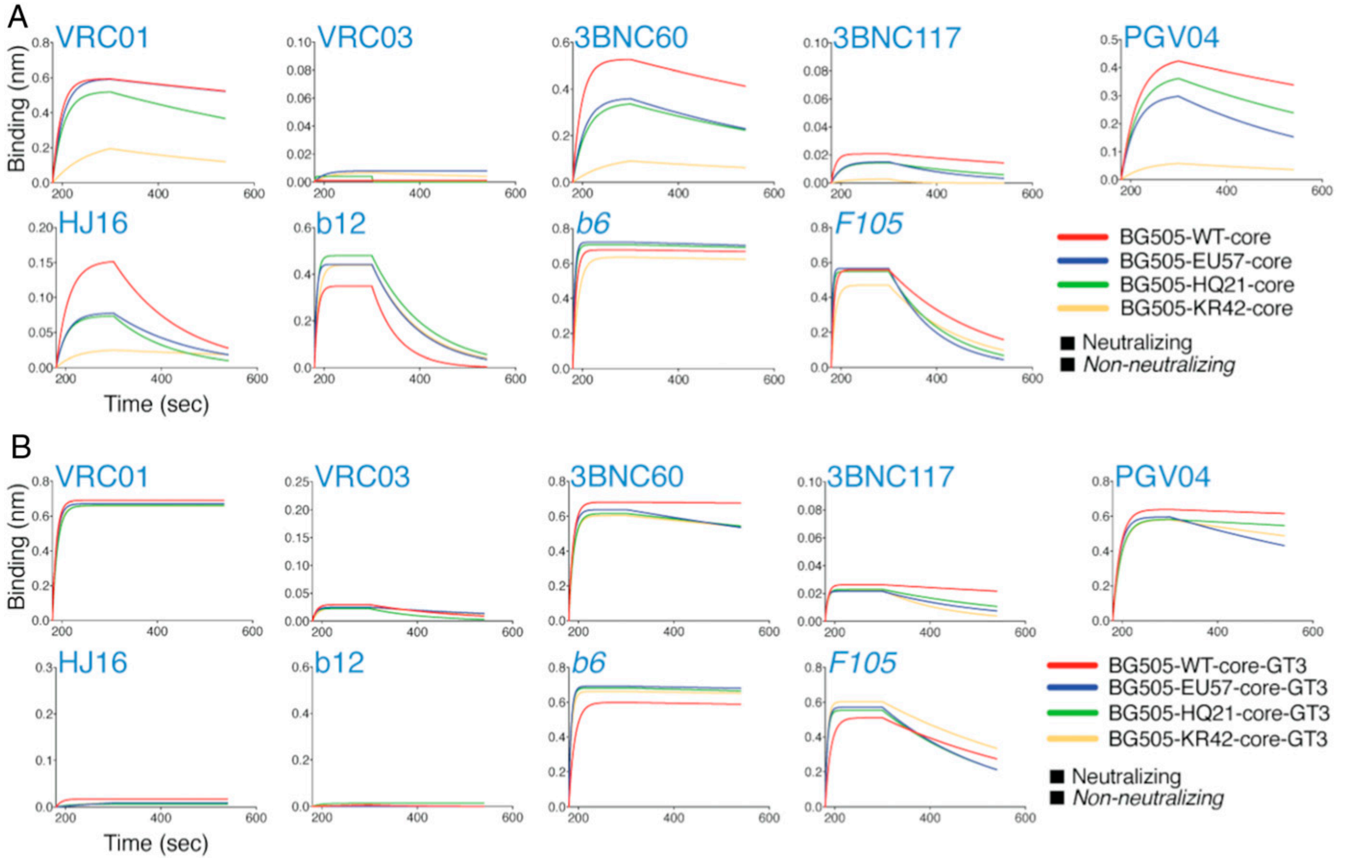

Fig. 6. BLI binding of HIV envelope CD4bs neutralizing (VRC01, VRC03, 3BNC60, 3BNC117, PGV04, HJ16, and b12) and nonneutralizing (b6 and F105) mAbs to (*A*) BG505/CD4bs variant cores or (*B*) core-GT3 versions. BLI binding curves (association, 120 s [180 to 300]; dissociation, 240 s [300 to 540]) are derived from binding responses of mAbs to soluble core and core-GT3 proteins. The binding response is shown in nanometers.

that are likely to be found in natural HIV infections but are also dissimilar.

In this work, we decided to use, as an optimal vaccination panel, a set of three antigens with an average sequence distance of five residues among the selected residues. This is a reasonable but somewhat arbitrary choice, because there is not enough information to determine these parameters. It is known that too large a sequence distance is unfavorable and causes premature extinction of the maturing B cell lineages (23). Further, the maximum allowed sequence distance is correlated with the number of antigens in the panels, and an optimal combination is possible to maximize the breadth of the produced antibodies (28). This optimal combination also depends on whether the antigens are administered as a mixture or in sequence. The values used in this work are reasonable guesses, given the lack of more accurate experimental or computational information. Using less than three antigens will probably fail to generate bnAbs (not much variability), while using a high number of antigens poses more experimental as well as, eventually, production issues. Other panel compositions will be explored, given that the Pareto analysis can be rerun in a straightforward manner. Moreover, the focus of this work is on the CD4bs, but the same analysis could be performed on other sites that contain conserved functional regions.

One optimized panel was chosen for experimental testing. The identified mutations were introduced intro three antigenic constructs: the trimeric BG505 SOSIP, the core, and core-GT3. The first is a native-like trimer, which can be used in the last step of a vaccination protocol. The core and core-GT3 are intermediate antigens, which can be recognized by antibodies elicited in the priming step and can act as a bridge between the primer and the native trimers. Experimental testing showed that all constructs, except the BG-KR trimer, can be synthesized and are stable in a native-like form. Moreover, they were shown to be recognized, at different levels, by known mAbs. These results strongly suggest that these constructs can be used in vaccination protocols. Further computational and experimental studies on these constructs are ongoing. Specifically, the next step is to test their efficacy in producing bnAbs in relevant mouse models.

Methods

Analysis of bnAbs Crystallographic Structures. The RCSB PDB (39) was explored to find crystallographic structures of broadly neutralized antibodies bound to HIV Env. The PDB was searched both directly and through the information available in the bnAber (40) and CATNAP (41) databases of bnAbs. Forty-two crystallographic structures of bnAbs bound to the CD4bs of the gp120 Env protein were extracted and analyzed; see list in *SI Appendix*. All of them share the same binding pose, except the bnAbs b12, CH103, VRC13, VRC16, and HJ16, whose binding poses appear shifted or rotated with respect to the others. These five structures were disregarded, and the analysis focused on the remaining 37. The main quantity of interest was which residues of gp120 are used in binding and are thus in contact with the bnAbs in the crystallographic structures. Two residues were defined to be in contact if their alpha carbons were separated by 7 Å or less. We determined which residues are most frequently used in binding, defining a usage scale going from 0% (never used) to 100% (used in all 37 structures).

Fitness Landscape and Escape Cost. The HIV Env fitness landscape has been recently inferred from gp160 sequence data (29). The model has the form

$$p(\mathbf{x}) = \frac{\exp[-E(\mathbf{x})]}{\sum_{\mathbf{x}'}\exp[-E(\mathbf{x})]}, \qquad [1]$$

where the amino acid sequence $\mathbf{x} = [x_1, x_2, \ldots, x_L]$ is a vector of length L in which each element x_i is the amino acid identity of residue i, and $p(\mathbf{x})$ is the predicted prevalence of sequence $\mathbf{x}$ in the viral population. The model depends on the sequence "energy" $E(\mathbf{x})$, which is a function of a set of fields $h_i(x_i)$ and couplings $J_{ij}(x_i, x_j)$ that depend on the identities of the amino acid at each position in the sequence as

$$E(\mathbf{x}) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L}\sum_{j=i+1}^{L} J_{ij}(x_i, x_j). \qquad [2]$$

According to this model, virus strains with a low (high) energy $E(\mathbf{x})$ are predicted to have high (low) prevalence and fitness $p(\mathbf{x})$. Furthermore, for various HIV proteins, it has been demonstrated that the rank order of intrinsic viral fitness of strains is statistically similar to the prevalence (42–45).

The strain energy defined above can be used to quantify the difficulty for a virus to introduce certain mutations, for example, to evade an immune response. This "escape cost" is represented by the difference in energy between the original sequence (with consensus amino acid c_i at residue i) and a sequence with a mutation at residue i to amino acid α,

$$\Delta E_i(a) = \sum_{\substack{x \\ x_i = c_i}} [E(\mathbf{x}) - E(\mathbf{x}')]p(\mathbf{x}). \qquad [3]$$

The summation is over all possible sequences, having the consensus amino acid at site$i$, and can be estimated by obtaining an ensemble of sequences using Markov chain Monte Carlo (29). This averaging over sequence backgrounds accounts for the fact that escape costs may depend on the network of couplings between particular amino acid mutations in the surrounding sequence. The above escape cost is an amino acid-specific measure; in order to obtain a residue-level measure of escape cost, this quantity is averaged over all possible mutations to a nonconsensus residue with (29)

$$\overline{\Delta E_i} = \frac{\sum_{a, a \neq c_i} \Delta E_i(a)\exp[-\Delta E_i(a)]}{\sum_{a, a \neq c_i} \exp[-\Delta E_i(a)]}. \qquad [4]$$

We carried out the averages in Eqs. 3 and 4 for the following reason. The goal of our work is to design an immunogen that will elicit an antibody response that can protect people from diverse strains of the infecting virus. Importantly, we are not trying to design a personalized vaccine for individuals, but for the population. The viruses that infect different people will have different sequence backgrounds (represented by the sequence ensemble $p(\mathbf{x})$). Thus, to find the residues of interest with the lowest or highest fitness costs associated with mutation, we average over the ensemble of possible sequence backgrounds. Similarly, in Eq. **4**, we average over all possible amino acids because different ones may have different fitness costs, and it is not a priori clear which one may evolve in different persons. Thus, the averaging procedures result in a more robust immunogen design for the population.

The quantity $\overline{\Delta E_i}$ is referred to in this manuscript simply as the "escape cost" for residue i. This quantity can also be considered a measure of variability of a residue; that is, low escape cost residues tend to have high variability, while high escape cost residues are generally conserved. However, this measure extends beyond simple amino acid conservation, as it accounts for the network of coupling interactions across the protein.

Antibody Footprint on the CD4bs. Atomistic models of the three complexes involving the VRC01 (8), VRC01GL (17), and DRVIA7 (31) bound to the BG505 SOSIP were created using, as a template, the crystallographic structure of BG505 SOSIP in complex with the scFv NIH45-46 antibody in the CD4bs (PDB ID code 5D9Q) (19). The VRC01 and NIH45-46 bnAbs share the same binding pose into the CD4bs; thus the use of PDB ID code 5D9Q as template should not introduce significant modeling errors. The three models were created using the software Modeler (46, 47), limiting the structure to one monomer of the SOSIP trimer and the variable region of the antibody Fab. Each complex was simulated using molecular dynamics for 10 ns, giving an ensemble of conformations to use for further analysis. Analyzing an ensemble of structures is needed to sample different conformations of the amino acid sidechains, and to possibly reduce modeling errors. The CHARMM36 empirical force field (48) was used as a model of the interatomic interactions, and the dynamics were integrated using OpenMM (49). The effect of solvent was modeled using the OBC2 implicit solvent model (50). No sugars were added to the models, thus simulating a fully deglycosylated complex. From the ensemble of conformations, the residues at the antigen/antibody interface were defined as those that had a large change in their solvent-accessible surface area (SASA) when the antibody was separated from the antigen. A residue was defined to be in contact and important in the binding if a difference in SASA was measured in more than 50% of the analyzed conformations in at least one of the studied complexes.

Pareto Frontier Analysis. Given a set $\mathcal{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_\mathrm{n}\}$ of n measurements of dimension d, the Pareto frontier $\mathcal{P}$ is defined as the subset of $\mathcal{M}$ for which there exists no other measurement more optimal in all d dimensions,

$$\mathcal{P} = \{\mathbf{m} \in \mathcal{M} : \{\mathbf{m}' \in \mathcal{M} : \mathbf{m}' < \mathbf{m}, \mathbf{m}' \neq \mathbf{m}\} = \varnothing\}, \qquad [5]$$

where $\mathbf{m}' < \mathbf{m}$ means that each element of $\mathbf{m}'$ is less than or equal to the corresponding element of $\mathbf{m}$, with at least one strict inequality (i.e., $\mathbf{m}'$ is more optimal than $\mathbf{m}$). The Pareto frontier is continually updated as new measurements are taken, according to the algorithm

```
▷ Initialize an empty Pareto frontier
$\mathcal{P} = \varnothing$
▷ Add initial measurement, $\mathbf{m}_1$, to Pareto frontier
$\mathcal{P} \leftarrow \{\mathbf{m}_1\}$
▷ Take measurements 2 through $N$
FOR $\mathbf{m} = \mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_N$
    ▷ Test if measurement $\mathbf{m}$ falls on Pareto frontier
    IF $\{\mathbf{p} \in \mathcal{P} : \mathbf{p} < \mathbf{m}, \mathbf{p} \neq \mathbf{m}\} = \varnothing$
        ▷ Add $\mathbf{m}$ to Pareto frontier
        $\mathcal{P} \leftarrow \mathcal{P} \cup \{\mathbf{m}\}$
        ▷ Remove elements no longer on Pareto frontier
        $\mathcal{P} \leftarrow \{\mathbf{p} \in \mathcal{P} : \{\mathbf{p}' \in \mathcal{P} : \mathbf{p}' < \mathbf{p}, \mathbf{p}' \neq \boldsymbol{p}\} = \varnothing\}$
    END IF
END FOR.
```

This analysis was used to select, from a pool of randomly generated immunogens panels, the ones minimizing the sequence similarity and energy, as discussed in the *Results and Discussion*. To assess convergence of the Pareto front with the number of panels tested, the "Pareto length" $L_\mathcal{P}$ was monitored, which was defined as

$$L_\mathcal{P} = \sum_{i=1}^{|\mathcal{P}|-1} \left\| \mathbf{p}_{(i)} - \mathbf{p}_{(i+1)} \right\|_2. \qquad [6]$$

$\mathbf{p}_{(i)}$ with $i = 1,2, \ldots, |\mathcal{P}|$ are the order statistics of the panels on the Pareto front (i.e., ordered by increasing energy). In other words, the Pareto length is the sum of Euclidean distances between all adjacent pairs of panels on the Pareto front. The average of the panels on the evolving Pareto front was also evaluated as

$$\overline{p} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \frac{\mathbf{p} - \mathbf{p}_{(0)}}{\mathbf{p}_{(|\mathcal{P}|)} - \mathbf{p}_{(0)}}. \qquad [7]$$

These metrics are plotted in Fig. 7 for increasing numbers of panels tested.

Design, Expression, and Purification of Proteins. To express soluble trimers, SOSIP.664 HIV Env trimer modification were incorporated into Env-encoding sequence of previously described BG505 isolate (51) and the BG505 CD4bs variants based on Env sequences corresponding to GenBank accession numbers (EU577271, HQ217523, and KR423280). Briefly, the following modifications were incorporated into these Envs for soluble trimer expression: 1) The Env leader sequence was replaced by Tissue Plasminogen Activator signal sequence for higher protein expression; 2) a disulfide bond was introduced between gp120 and gp41 subunits by substituting residues A501-C and T605-C, respectively, in gp120 and gp41; 3) the gp120 REKR cleavage site was replaced by a furin R6 site (RRRRRR) for enhancing cleavage efficiency between gp120 and gp41; and 4) an I559P substitution in gp41 to stabilize the soluble trimer protein. In addition, a stop codon was added to the gp41ECTO C terminus at HXB2 residue 664 position to truncate the protein for soluble expression. The core and core-GT3 versions of BG505 and its CD4bs variants were constructed by employing design strategies described elsewhere (24). The codon-optimized gene constructs were synthesized (Geneart, Life Technologies) and cloned into the phCMV3 vector (Genlantis). Recombinant Env proteins were expressed in HEK293F cells as described previously (51). Briefly, the SOSIP.664 trimer (cotransfected with furin; 2:1 Env:furin ratio)

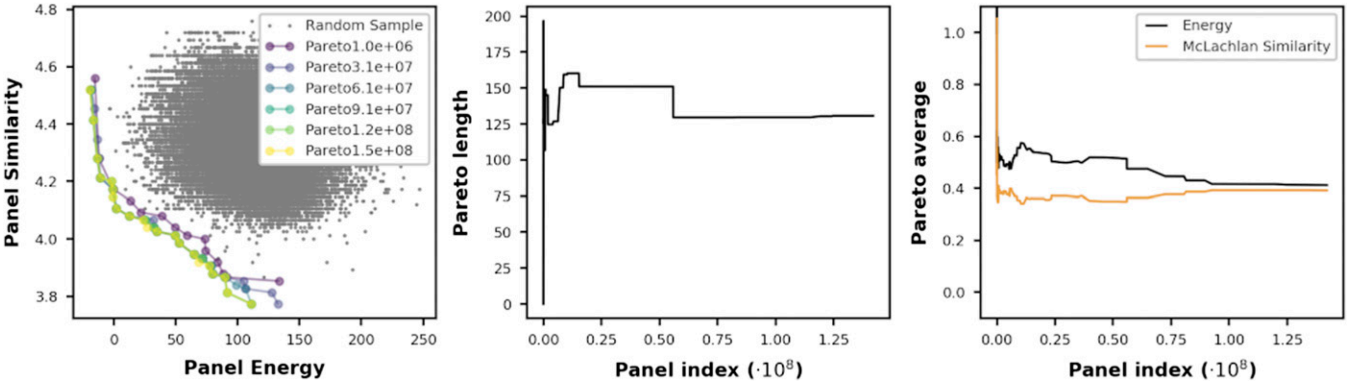

Fig. 7. Convergence of the Pareto frontier.

and core/core-GT3 gene encoding plasmids were transfected into HEK293F cells using PEI-MAX 4000 transfection reagent (Polysciences, Inc.). The secreted soluble proteins were purified from cell supernatants after 5 d using agarose-bound Gallanthus Nivalis Lectin (Vector Labs). The affinity-purified proteins were SEC-separated with a Superdex 200 10/300 GL column (GE Healthcare) in phosphate-buffered saline with tryptic soy broth.

BLI Binding Assay. The BLI binding experiments of Abs to the proteins were performed with an Octet K2 system (ForteBio, Pall Life Sciences) as described previously (52). Briefly, HIV Env-specific mAbs (10 μg/mL in phosphate buffered saline with Tween 20 [PBST]) were immobilized onto hydrated anti-human IgG-Fc biosensors (AHC: ForteBio) for 60 s to achieve a binding response of at least 1.0. Following the Ab capture step, the sensor was placed in a PBST wash buffer to remove the nonspecifically bound Ab and establish a baseline signal. Next, the antibody-immobilized sensor was dipped into a solution containing soluble protein as analyte and incubated for 120 s at 1,000 rpm. This step denotes association binding. Following this, the protein bound to antibody-immobilized sensor was removed from the analyte solution and placed into the PBST buffer for 240 s at 1,000 rpm. This step denotes the dissociation binding. The sensograms were corrected with the blank reference and fit (1:1 binding kinetics model) with the ForteBio Data Analysis version 9 software using the global fitting function.

Data Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. S.C., A.K.C., and M.K. received financial support from the Lawrence Livermore National Laboratory under Grant 17-ERD-043 (LLC Award B620960) and from the CHARMM Development Project. This work was supported by the Consortium for HIV/AIDS Vaccine Development (CHAVD) NIH Award UM1 AI44462 (to D.R.B), the IAVI Neutralizing Antibody Center, and the Bill and Melinda Gates Foundation (Award OPP1196345 to D.R.B.). We are grateful to Victor Ovchinnikov for helpful discussions and comments on the manuscript.

1. A. S. Fauci, An HIV vaccine is essential for ending the HIV/AIDS pandemic. *JAMA* **318**, 1535–1536 (2017).
2. D. R. Burton, L. Hangartner, Broadly neutralizing antibodies to HIV and their role in vaccine design. *Annu. Rev. Immunol.* **34**, 635–659 (2016).
3. D. Corti, A. Lanzavecchia, Broadly neutralizing antiviral antibodies. *Annu. Rev. Immunol.* **31**, 705–742 (2013).
4. D. R. Burton *et al*., Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* **266**, 1024–1027 (1994).
5. J. Huang *et al*., Identification of a CD4-binding-site antibody to HIV that evolved near-pan neutralization breadth. *Immunity* **45**, 1108–1121 (2016).
6. Y. Li *et al*., Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nat. Med.* **13**, 1032–1034 (2007).
7. J. F. Scheid *et al*., Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633–1637 (2011).
8. T. Zhou *et al*., Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**, 811–817 (2010).
9. F. Gao *et al*., Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group M consensus envelope glycoprotein. *J. Virol.* **79**, 1154–1163 (2005).
10. B. Gaschen *et al*., Diversity considerations in HIV-1 vaccine selection. *Science* **296**, 2354–2360 (2002).
11. D. H. Barouch *et al*., Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nat. Med.* **16**, 319–323 (2010).
12. M. Bonsignori *et al*., Antibody-virus co-evolution in HIV infection: Paths for HIV vaccine development. *Immunol. Rev.* **275**, 145–160 (2017).
13. B. T. Korber, N. L. Letvin, B. F. Haynes, T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. *J. Virol.* **83**, 8300–8314 (2009).
14. F. Klein *et al*., Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* **153**, 126–138 (2013).
15. V. Ovchinnikov, J. E. Louveau, J. P. Barton, M. Karplus, A. K. Chakraborty, Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies. *eLife* **7**, e33038 (2018).
16. A. Escolano *et al*., Sequential immunization elicits broadly neutralizing anti-HIV-1 antibodies in Ig Knockin mice. *Cell* **166**, 1445–1458.e12 (2016).
17. J. Jardine *et al*., Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340**, 711–716 (2013).
18. L. Scharf *et al*., Structural basis for HIV-1 gp120 recognition by a germ-line version of a broadly neutralizing antibody. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6049–6054 (2013).
19. J. G. Jardine *et al*., Minimally mutated HIV-1 broadly neutralizing antibodies to guide reductionist vaccine design. *PLoS Pathog.* **12**, e1005815 (2016).
20. T. Zhou *et al*.; NISC Comparative Sequencing Program, Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258 (2013).
21. J. G. Jardine *et al*., HIV-1 VACCINES. Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen. *Science* **349**, 156–161 (2015).
22. P. D. Kwong, J. R. Mascola, G. J. Nabel, Broadly neutralizing antibodies and the search for an HIV-1 vaccine: The end of the beginning. *Nat. Rev. Immunol.* **13**, 693–701 (2013).
23. S. Wang *et al*., Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* **160**, 785–797 (2015).
24. B. Briney *et al*., Tailored immunogens direct affinity maturation toward HIV neutralizing antibodies. *Cell* **166**, 1459–1470.e11 (2016).
25. P. Dosenovic *et al*., Immunization for HIV-1 broadly neutralizing antibodies in human Ig Knockin mice. *Cell* **161**, 1505–1515 (2015).
26. J. M. Steichen *et al*., HIV vaccine design to target germline precursors of glycan-dependent broadly neutralizing antibodies. *Immunity* **45**, 483–496 (2016).
27. M. Tian *et al*., Induction of HIV neutralizing antibody lineages in mice with diverse precursor repertoires. *Cell* **166**, 1471–1484.e18 (2016).
28. J. S. Shaffer, P. L. Moore, M. Kardar, A. K. Chakraborty, Optimal immunization cocktails can promote induction of broadly neutralizing Abs against highly mutable pathogens. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7039–E7048 (2016).
29. R. H. Y. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, M. R. McKay, Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E564–E573 (2018).
30. P. D. Kwong, J. R. Mascola, Human antibodies that neutralize HIV-1: Identification, structures, and B cell ontogenies. *Immunity* **37**, 412–425 (2012).
31. L. Kong *et al*., Key gp120 glycans pose roadblocks to the rapid development of VRC01-class antibodies in an HIV-1-infected Chinese donor. *Immunity* **44**, 939–950 (2016).
32. A. D. McLachlan, Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **64**, 417–437 (1972).
33. A. Jahan, K. L. Edwards, M. Bahraminasab, *Multi-criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design* (Butterworth-Heinemann, 2016).
34. J.-P. Julien *et al*., Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1477–1483 (2013).
35. D. Lyumkis *et al*., Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1484–1490 (2013).
36. R. W. Sanders *et al*., HIV-1 VACCINES. HIV-1 neutralizing antibodies induced by native-like envelope trimers. *Science* **349**, aac4223 (2015).

37. L. E. McCoy *et al*., Holes in the glycan shield of the native HIV envelope are a target of trimer-elicited neutralizing antibodies. *Cell Rep.* **16**, 2327–2338 (2016).
38. P. Pugach *et al*., A native-like SOSIP.664 trimer based on an HIV-1 subtype B env gene. *J. Virol.* **89**, 3380–3395 (2015).
39. H. M. Berman *et al*., The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
40. A. M. Eroshkin *et al*., bNAber: Database of broadly neutralizing HIV antibodies. *Nucleic Acids Res.* **42**, D1133–D1139 (2014).
41. H. Yoon *et al*., CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res.* **43**, W213–W219 (2015).
42. J. P. Barton *et al*., Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat. Commun.* **7**, 11660 (2016).
43. A. L. Ferguson *et al*., Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
44. J. K. Mann *et al*., The fitness landscape of HIV-1 gag: Advanced modeling approaches and validation of model predictions by in vitro testing. *PLOS Comput. Biol.* **10**, e1003776 (2014).
45. K. Shekhar *et al*., Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **88**, 062705 (2013).
46. A. Šali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
47. B. Webb, A. Sali, Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* **54**, 5.6.1–5.6.37 (2002).
48. R. B. Best *et al*., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
49. P. Eastman *et al*., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).
50. A. Onufriev, D. Bashford, D. A. Case, Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383–394 (2004).
51. R. W. Sanders *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).
52. R. Andrabi *et al*., Glycans function as anchors for antibodies and help drive HIV broadly neutralizing antibody development. *Immunity* **47**, 524–537.e3 (2017).