Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies

Raymond H. Y. Louie[a,b,1], Kevin J. Kaczorowski[c,d,1], John P. Barton[d,e,f], Arup K. Chakraborty[c,d,e,f,g,2], and Matthew R. McKay[a,h,2]

[a]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong; [b]Institute for Advanced Study, Hong Kong University of Science and Technology, Kowloon, Hong Kong; [c]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [e]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; [f]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA 02139; [g]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139; and [h]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong

Contributed by Arup K. Chakraborty, December 11, 2017 (sent for review October 10, 2017; reviewed by Pamela J. Bjorkman and Martin Weigt)

HIV is a highly mutable virus, and over 30 years after its discovery, a vaccine or cure is still not available. The isolation of broadly neutralizing antibodies (bnAbs) from HIV-infected patients has led to renewed hope for a prophylactic vaccine capable of combating the scourge of HIV. A major challenge is the design of immunogens and vaccination protocols that can elicit bnAbs that target regions of the virus's spike proteins where the likelihood of mutational escape is low due to the high fitness cost of mutations. Related challenges include the choice of combinations of bnAbs for therapy. An accurate representation of viral fitness as a function of its protein sequences (a fitness landscape), with explicit accounting of the effects of coupling between mutations, could help address these challenges. We describe a computational approach that has allowed us to infer a fitness landscape for gp160, the HIV polyprotein that comprises the viral spike that is targeted by antibodies. We validate the inferred landscape through comparisons with experimental fitness measurements, and various other metrics. We show that an effective antibody that prevents immune escape must selectively bind to high escape cost residues that are surrounded by those where mutations incur a low fitness cost, motivating future applications of our landscape for immunogen design.

HIV | fitness landscape | envelope protein | statistical inference | broadly neutralizing antibodies

After over three decades of effort, there is now renewed hope for an effective vaccine against HIV. This is because of the isolation of broadly neutralizing antibodies (bnAbs) that are capable of neutralizing diverse HIV strains in vitro and progress made toward inducing them by vaccination (1–7), and reports of protection against infection for macaques immunized with a T cell-based vaccine (8, 9). However, much progress still needs to be made to realize the goal of deploying an effective vaccine, and also to develop rational strategies for optimizing immunogens and vaccination protocols. A related issue is the choice of combinations of bnAbs for passive therapy of infected persons, an approach that has shown promise in humans and macaques (10–12).

Because HIV is highly mutable, one key challenge is that the virus can evolve mutations that abrogate the binding of antibodies to HIV's envelope proteins or T cell receptors to peptides derived from viral proteins, while also preserving fitness (the ability to properly assemble the virus, replicate, and propagate infection). This allows HIV to escape from human immune responses. On the other hand, some parts of the viral proteome are vulnerable to mutations because they result in a severe loss of fitness. Vaccination strategies aimed at targeting these parts of the proteome and not the ones that readily allow escape might be effective. The design of effective vaccination strategies would therefore benefit from knowledge of the fitness landscape of viral proteins—that is, knowledge of the fitness of the virus as a function of its amino acid sequence.

A conventional approach to estimate the fitness landscape of a virus is to assume that the fitness of a sequence is directly related to the extent to which the amino acids at a residue are conserved in circulating viral strains; more conserved residues are ones where mutations are predicted to result in a larger fitness penalty. This approach was used to obtain a landscape experimentally for the HIV envelope protein, gp160 (13), and computationally for other HIV proteins (14, 15). However, it ignores epistatic coupling between mutations at different residues, which is known to be an important factor for viral fitness (16–18). For example, if a mutation that allows the virus to evade an immune response occurs at a particular relatively conserved residue, it is likely to significantly impair viral fitness. However, if another mutation at a different residue has a compensatory effect, the fitness cost incurred by making the primary immune-evading mutation can be partially restored. Because HIV is highly mutable and also exhibits a high replication rate, such compensatory mutations can be accessed in vivo, and have been found to be a significant factor in determining viral fitness and mutational escape pathways (16–19). Thus, to define the mutational vulnerabilities of a virus like HIV in vivo, and thus guide the design of vaccination strategies, the desired fitness landscape must include the effects

Significance

An effective vaccine for HIV is still not available, although recent hope has emerged through the discovery of antibodies capable of neutralizing diverse HIV strains. Nonetheless, there exist mutational pathways through which HIV can evade known broadly neutralizing antibody responses. An ideal vaccine would elicit broadly neutralizing antibodies that target parts of the virus's spike proteins where mutations severely compromise the virus's fitness. Here, we employ a computational approach that allows estimation of the fitness landscape (fitness as a function of sequence) of the polyprotein that comprises HIV's spike. We validate the inferred landscape through comparisons with diverse experimental measurements. The availability of this fitness landscape will aid the rational design of immunogens for effective vaccines.

Author contributions: R.H.Y.L., K.J.K., A.K.C., and M.R.M. designed research; R.H.Y.L., K.J.K., A.K.C., and M.R.M. performed research; R.H.Y.L., K.J.K., J.P.B., A.K.C., and M.R.M. analyzed data; and R.H.Y.L., K.J.K., J.P.B., A.K.C., and M.R.M. wrote the paper.

Reviewers: P.J.B., California Institute of Technology; and M.W., Université Pierre et Marie Curie.

The authors declare no conflict of interest.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]R.H.Y.L. and K.J.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: arupc@mit.edu or m.mckay@ust.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717765115/-/DCSupplemental.

of coupling between mutations at different residues. Unfortunately, despite continuing progress (13, 20, 21), experimentally obtaining a fitness landscape that includes coupling information for the majority of proteins in HIV is infeasible, because the number of experimental fitness values that need to be measured is prohibitively large. For example, there are on the order of 10^7 experimental fitness values that would need to be measured for gp160 (Fig. 1), the polyprotein that comprises the HIV spike that is targeted by antibodies.

One approach that has been successful in obtaining the fitness landscape of HIV proteins, including the effects of couplings, has relied on inferring this information from the sequences of circulating strains [a multiple sequence alignment (MSA)] (19, 22–25). First, a maximum-entropy–based computational approach (*SI Appendix*, *Text 2.1*) was used to infer a "prevalence landscape" from the sequence data, taking into account couplings between residues; the prevalence landscape describes the probability of observing a virus with particular protein sequence in circulation (Fig. 2*A*). Theoretical analyses and tests against in vitro and clinical data (19, 22, 26) suggest that, because of the evolutionary history of HIV and the diversity of (largely ineffective) immune responses that the HIV population has been subjected to, the prevalence landscape is a reasonably good proxy for the fitness landscape. This approach has been applied and validated for various internal proteins of HIV (19, 22–25). Similar methods have also been employed to predict the fitness effects of mutations in bacterial proteins (27). However, a fitness landscape for gp160 has not been previously obtained, despite its importance as a target of antibody responses. This is because the inference problem is far more challenging. In particular, the gp160 primary sequence is more than twice as long as any other HIV protein, and it is also among the most variable. This leads to an explosion in the number of model parameters to be estimated (Fig. 1). Standard inference approaches, such as those based on gradient descent algorithms [commonly referred to as Boltzmann machine-learning (BML) methods] or Markov chain Monte Carlo-based simulation methods (22, 28) are intractable for the gp160 protein.

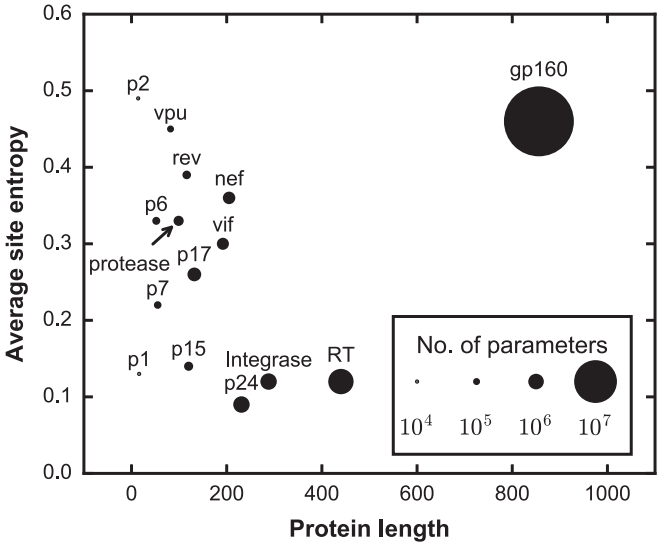

Fig. 1. For different HIV proteins, the number of parameters, average site entropy, and protein length are shown. gp160 has orders of magnitude more parameters than the majority of the other HIV proteins, the longest length, and one of the largest average site entropy. Note that, to obtain a landscape purely based on experiments, the number of required in vitro experimental fitness values is proportional to the number of parameters. The amino acid MSA for each protein was downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database (https://www.hiv.lanl.gov/).

To address this issue, we introduce a computational framework (Fig. 2*B*) that efficiently and accurately calculates the parameters of a maximum-entropy model for the gp160 fitness landscape. We validate the inferred fitness landscape by testing predictions of fitness against in vitro measurements of the fitness of nearly 100 HIV strains bearing mutations in gp160, testing predictions of protein contacts against structural data, and showing that our fitness landscape is consistent with known escape mutations. The fitness landscape that we report can help guide the design of immunogens aimed toward eliciting bnAbs, the choice of bnAb combinations for passive therapy, and the choice of immunogens for the T cell component of a vaccine.

Computational Method

The fitness landscape is specified as a probabilistic model, chosen as the model having the maximum entropy (i.e., the model least biased by intuition), subject to the constraint of reproducing the single- and double-mutant probabilities observed in the MSA (*SI Appendix*, Eq. **S5**). For a given length-L amino acid sequence $\mathbf{x}=[x_1,x_2,\ldots,x_L]$, where x_i can take on any of the 20 naturally occurring amino acids or gap in the MSA, this maximum-entropy model assigns the probability (22):

$$p_{\mathbf{h},\mathbf{J}}(\mathbf{x})=\frac{\exp[-E(\mathbf{x})]}{\sum_{\mathbf{x}'}\exp[-E(\mathbf{x}')]},\ E(\mathbf{x})=\sum_{i=1}^{L}h_i(x_i)+\sum_{i=1}^{L}\sum_{j=i+1}^{L}J_{ij}(x_i,x_j), \quad [1]$$

where, in analogy with statistical mechanics, we refer to $E(\mathbf{x})$ as the energy of sequence $\mathbf{x}$. The parameters $h_i(x_i)$ and $J_{ij}(x_i,x_j)$, referred to as fields and couplings, respectively, can be obtained by solving the following convex optimization problem (*SI Appendix*, *Text 2.1*):

$$(\mathbf{h}^*,\mathbf{J}^*)=\arg\min_{\mathbf{h},\mathbf{J}}\mathrm{KL}\left(p_0\|\,p_{\mathbf{h},\mathbf{J}}\right), \quad [2]$$

subject to the single- and double-mutant probabilities matching those in the MSA (*SI Appendix*, Eq. **S5**). Here, p_0 denotes the patient-weighted sequence probability distribution from the MSA (*SI Appendix*, *Text 2.1*), and $\mathrm{KL}(p_0\|\,p_{\mathbf{h},\mathbf{J}})$ denotes the Kullback–Leibler divergence between p_0 and $p_{\mathbf{h},\mathbf{J}}$ (*SI Appendix*, Eq. **S7**). This problem can be solved in principle by BML algorithms; however, such algorithms require computing a gradient function that involves an exponential number of terms in L, thereby making direct computation infeasible for the gp160 protein. A solution to this challenge is provided for other HIV proteins by use of a cluster expansion approach along with BML (24, 29, 30). However, achieving reliable statistical inference for gp160 is difficult due to the large parameter space (Fig. 1) and the relatively small number of sequences available for gp160 (*SI Appendix*, *Text 1*). The method described here addresses these issues via three main steps (Fig. 2*B* and *SI Appendix*, *Text 2*). The most important of these is the introduction of a computationally efficient algorithm that produces accurate initial estimates of the fields and couplings. This method is augmented by the refinement of these estimates using BML, which, due to the accuracy of the initialization, converges quickly, and the application of a variable-combining method for reducing the effective number of parameters, which is more systematic than past such approaches.

Our approach for providing an initial estimate of the field and coupling parameters is based on the principle of minimum probability flow (MPF) (31), which was successfully applied to different applications such as deep belief networks and independent component analysis (31), although to our knowledge it has yet to be used for the analysis of protein sequence data. We first describe the general idea of MPF as applied to our maximum-entropy problem, and then discuss some distinctions with related work. The MPF principle leads to replacing $p_{\mathbf{h},\mathbf{J}}$ in

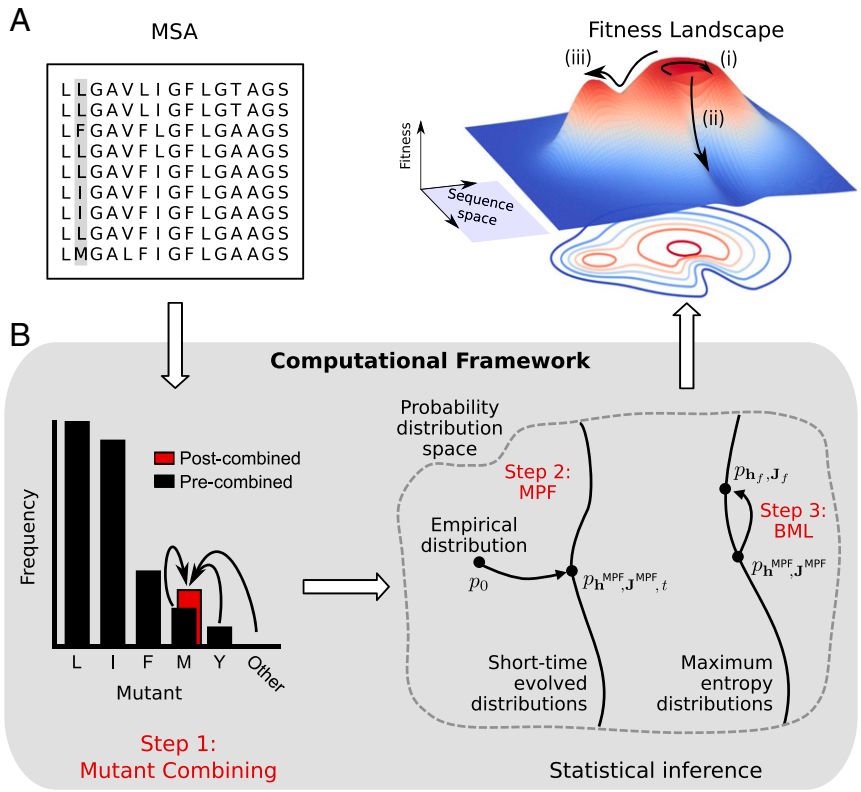

Fig. 2. (*A*) The general framework used for inferring a fitness landscape for gp160. Protein sequence data, in the form of a multiple sequence alignment (MSA), is used to infer a maximum-entropy, or least-biased, probability distribution of sequences, subject to recapitulating the observed one- and two-point mutation probabilities. This distribution is termed the sequence prevalence landscape (*SI Appendix, Text 2.1*). Based on past work on other HIV proteins, we assume that protein fitness is directly proportional to sequence prevalence. As shown in a schematic depiction (*Right*), the fitness landscape traces out fitness values over the space of possible protein sequences and highlights several features expected of the gp160 landscape: HIV often escapes even in the presence of bnAbs, suggesting that many mutational pathways may have small fitness costs (*i*); other escape pathways may lead to a large decrease in fitness (*ii*); but compensatory mutations can restore fitness (*iii*). (*B*) Our computational framework includes three steps to solve the maximum-entropy problem. Step 1 involves reducing the number of mutants from the original MSA (*SI Appendix, Text 2.3*). The bar chart shows the single-mutant probabilities for one particular residue (highlighted in gray in the MSA depicted in *A*). Infrequent amino acids are combined together and treated as a single mutant ("post-combined"). Step 2 involves solving a modified minimum probability flow (MPF) objective function involving the KL divergence between the empirical distribution and a short-time evolved distribution (Eq. **3**), to include regularization and arbitrary number of mutants (*SI Appendix, Text 2.3*). Step 3 involves refining the parameters from step 2 using a Boltzmann machine-learning (BML) algorithm (*SI Appendix, Text 2.3*), which produces the field ($\mathbf{h}_f$) and coupling ($\mathbf{J}_f$) parameters of the landscape.

the optimization problem (Eq. **2**) with an alternate (related) distribution that is simpler to optimize, and which is expected to yield an accurate approximation to the desired maximum-entropy solution. Specifically, the maximum-entropy form (Eq. **1**) is viewed as the equilibrium distribution of a Markov process with appropriately specified deterministic dynamics (*SI Appendix, Text 2.3*) that evolves the empirical data distribution p_0 toward $p_{\mathbf{h},\mathbf{J}}$. Based on these dynamics, the original optimization problem (Eq. **2**) is replaced with the alternative problem:

$$\left(\mathbf{h}^{\mathrm{MPF}},\mathbf{J}^{\mathrm{MPF}}\right)=\arg\min_{\mathbf{h},\mathbf{J}}\mathrm{KL}\left(p_0\|\,p_{\mathbf{h},\mathbf{J};t}\right), \qquad [3]$$

where $p_{\mathbf{h},\mathbf{J};t}$ (*SI Appendix*, Eq. **S13**) represents the distribution obtained by running the dynamics for time t. While this distribution, as well as the associated KL divergence in Eq. **3**, is still quite complicated, it is greatly simplified for small values of t. Specifically, for small t, the KL divergence is well approximated by a linear function of t (*SI Appendix, Text 2.3*):

$$\mathrm{KL}\left(p_0\|\,p_{\mathbf{h},\mathbf{J};t}\right)\approx t\mathrm{K}_{\mathbf{h},\mathbf{J}},$$

where $\mathrm{K}_{\mathbf{h},\mathbf{J}}$ is a simple function of the field and coupling parameters $\mathbf{h}$, $\mathbf{J}$ (see *SI Appendix*, Eq. **S16** for the explicit formula). The parameters that minimize this function can then be easily found using a standard gradient descent algorithm. Importantly, choosing small t to facilitate an expansion of the KL divergence (Eq. **3**) results in an efficient objective function $\mathrm{K}_{\mathbf{h},\mathbf{J}}$, which involves only a quadratic number of terms in L (*SI Appendix*, Eq. **S15**), as opposed to the original problem (Eq. **2**), which was exponential in L. The estimates obtained by minimizing $\mathrm{K}_{\mathbf{h},\mathbf{J}}$ are also statistically consistent (31): if the data are drawn from the maximum-entropy model (Eq. **1**), then the parameter estimates become arbitrarily accurate given a sufficiently large number of samples.

Of direct relevance to our problem, the MPF procedure was previously applied to learn Ising spin-glass models (31), which have the form of the maximum-entropy distribution (Eq. **1**), but restricted to modeling only two amino acids per residue. Our approach (*SI Appendix, Text 2.3*) extends that procedure to make it suitable for the gp160 problem. First, we allow for arbitrary numbers of amino acids per residue, which is essential due to the large sitewise amino acid variability observed in the MSA. Second, we introduce regularization to control sampling noise (*SI Appendix*, Eq. **S17** and *Text 2.3*), which is required due to the large number of parameters and limited sequence data for gp160. The regularization parameters are chosen such that the model parameters yield a suitable balance between statistical overfitting and underfitting, based on a previously defined metric

(32). This is achieved by appropriately biasing the estimated single- and double-mutant probabilities through regularization, to faithfully capture real mutational variability and ignore variability resulting from finite sampling (*SI Appendix*, Eq. **S18** and *Text 2.4*).

By solving a proxy optimization problem, the MPF inference procedure yields a set of inferred model parameters that serve as approximations to the desired optimal maximum-entropy parameters (Eq. **2**). These parameters are then refined using a gradient descent-based BML algorithm, referred to as RPROP (33), with an additional regularization term (*SI Appendix*, *Text 2.4*). While BML algorithms are intractable when applied to the gp160 maximum-entropy problem with arbitrary initialization (Eq. **2**), such algorithms can converge quickly when accurate initial parameters resulting from the MPF-based inference procedure are provided.

Multiple amino acids are observed at each residue of gp160 in the MSA. Here, we restrict the number of amino acids explicitly modeled in the inference procedure to decrease the computational burden (30). We propose a general data-driven approach to combine the least frequent mutants (nonconsensus amino acids) together. Combining such mutants also makes sense statistically since the low-frequency mutants are particularly sensitive to sampling noise, and their corresponding parameters are more challenging to meaningfully estimate. This approach, applicable to any protein, can substantially decrease the number of parameters to be estimated. To achieve this, as in the past (30), for residue i we model only the k_i most frequent mutants, while grouping the remaining $q_i - k_i + 1$ mutants together, where q_i denotes the number of mutants at residue i as observed in the MSA. Here, k_i is chosen such that, upon grouping, the entropy at residue i is at least a certain fraction ϕ of the corresponding entropy without grouping. The design question is to choose an appropriate value of ϕ. In previous work (19, 22, 30), this factor was chosen arbitrarily. However, for a protein like gp160, which has a significantly larger number of parameters than the other HIV proteins (Fig. 1), a more judicious choice of ϕ may have considerable benefits. A lower value will combine more mutants together, thus decreasing the number of parameters to be estimated and reducing the computational burden. However, combining too many mutants can result in a loss of useful information regarding amino acid identities. These quantities can be quantified in the form of a statistical bias (*SI Appendix*, *Text 2.2*).

To balance these competing issues systematically, we introduce a method for selecting ϕ based on the sequence data (*SI Appendix*, *Text 2.2*). It is described briefly as follows. For any given ϕ, for each residue i we can specify k_i as indicated above. Letting $f_i(a)$ denote the frequency of amino acid a at residue i, this leads to a modified (sparsified) model in which the frequencies of the modified model are now as follows:

$$\bar{f}_i(a) = \begin{cases} f_i(a) & \text{if } a < k_i + 1 \\ \bar{f}_i & \text{if } a = k_i + 1 \\ 0 & \text{if } a > k_i + 1 \end{cases}, \qquad [4]$$

where $\bar{f}_i = \sum_{a=k_i+1}^{q_i} f_i(a)$ represents the frequency of the coarse-grained amino acid. The squared error/bias of this model (i.e., for the specific ϕ) at residue i is then estimated as $\sum_{a=1}^{q_i} [f_i(a) - \bar{f}_i(a)]^2$, while the total variance of the amino acid frequencies at residue i is estimated as $\sum_{a=1}^{q_i} f_i(a)[1 - f_i(a)]/P$. Our aim is to select a level of coarse-graining that corresponds to the fractional entropy captured, ϕ, such that these quantities are commensurate. Intuitively, this approach seeks to select the sparsest model (i.e., choosing ϕ giving the most coarse-grained combining) such that the resulting errors caused by combining do not generally exceed the statistical fluctuations in the estimated amino acid frequencies from the MSA. A significant advantage of such variable selection approach is that it can substantially reduce the number of maximum-entropy parameters to infer (i.e., reducing the number of fields and couplings in Eq. **1**), thereby simplifying the parameter estimation problem, but without sacrificing the predictive power of the inferred model.

To this end, denoting the ratio of fractional error to variance in the amino acid frequencies in the data (see above) as $\beta_i(\phi)$ and letting $\langle \beta_i(\phi) \rangle$ be the average of $\beta_i(\phi)$ taken over all residues i, we select ϕ by numerically searching for a value that yields $\langle \beta_i(\phi) \rangle$ close to 1.

Results

We applied the computational framework to a MSA of HIV-1 clade B gp160 amino acid sequences downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database (https://www.hiv.lanl.gov/). The MSA was processed to ensure both sequence and residue quality (*SI Appendix*, *Text 2.1*), resulting in $L = 815$ residues and 20,043 sequences belonging to 1,918 patients. These sequences were highly variable, involving an average pairwise Hamming distance of 0.1824, normalized by L.

The Computational Method Is Fast and Accurately Captures the Observed Sequence Statistics. When applied to the MSA of gp160, our computational framework took only 2.5 d of CPU time (12 h for MPF, 2 d for convergence of the subsequent BML) on a 16-core node with 128-GB RAM and 2.7-GHz processing speed. This is significant given the large number of parameters that needed to be estimated (~4.4 million). The computational efficiency was aided by the initial variable selection phase, which reduced the number of parameters by a factor of 6. The inferred model accurately reproduced the single- and double-mutant probabilities observed in the MSA (Fig. 3 *A* and *B*), as required. The fast convergence of the BML algorithm in our framework was due both to the significant reduction in parameters achieved with the initial selection phase, as well as the high accuracy of the initialization parameters estimated by MPF (Fig. 3 *A* and *B*). Nonetheless, the BML refinement led to a better fit of the single- and double-mutant probabilities, while also yielding a clear improvement in terms of connected correlations (Fig. 3*C*). Finally, while the model was designed to explicitly reproduce single- and double-mutant probabilities, the inferred parameters also captured higher order statistics (Fig. 3*D*). This information was not reflected by the model produced by MPF, emphasizing the importance of the subsequent BML estimation phase.

Inferred Fitness Landscape Accurately Predicts in Vitro Replicative Fitness Measurements. To assess the ability of the inferred model to capture the fitness of different strains of gp160, we compared predictions based on this model to in vitro measurements of HIV fitness. We compiled 98 fitness measurements from in vitro experiments using competition (16, 17, 34, 35) or infectivity assays (18, 36). Strains used in these experiments were constructed by engineering single, double, or higher order mutations on a CCR5-tropic strain. However, there was no explicit consideration given to choosing higher order mutants that exhibit epistatic coupling, known to be important in vivo (*Discussion*). The metric of fitness in our model is the energy, E in Eq. **1**. Larger values of E correspond to less-fit strains. Based on the values of E computed using our fitness landscape, we predicted the relative fitness of the experimentally tested strains. We then determined the rank correlation [based on a weighted Spearman correlation measure (*SI Appendix*, Eq. **S19** and *Text 3*)] between these values of E and the measured fitness values (Fig. 4*A*). As anticipated, we found a strong weighted negative correlation (*SI Appendix*, Eq. **S19**), given by $\bar{\rho} = -0.74$ between model energies and viral fitness (see *SI Appendix*, Fig. S1 and *Text 3* for results for each individual experiment). These results demonstrate the ability of the inferred landscape to discriminate the relative intrinsic fitness of different HIV strains with mutations in gp160.

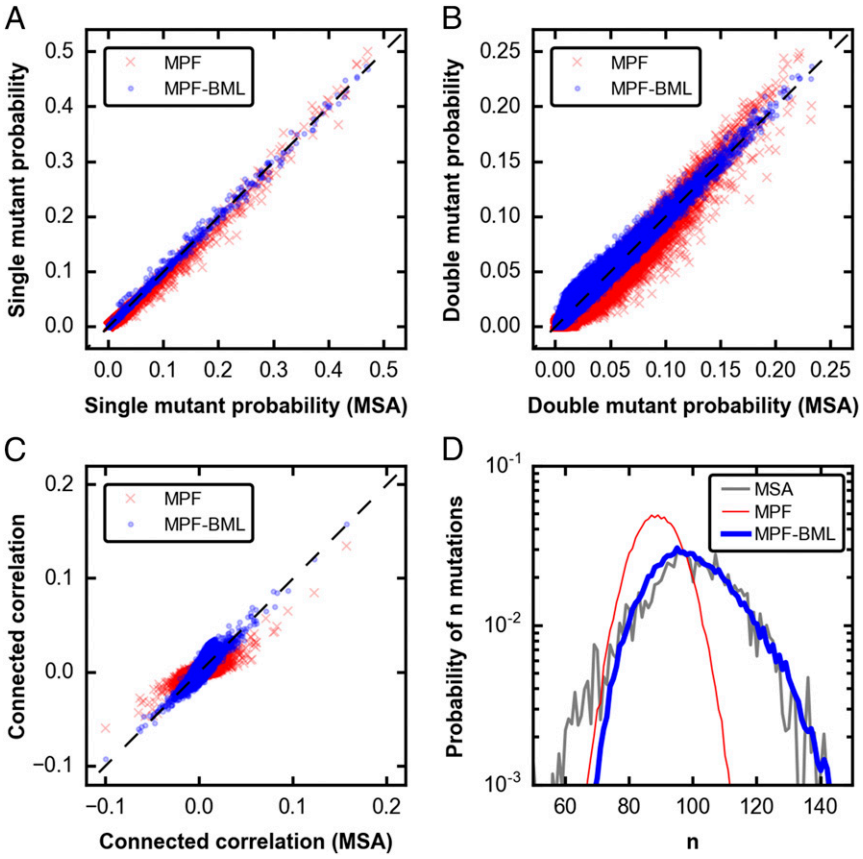

Fig. 3. (*A* and *B*) Scatter plot of the MPF/MPF-BML vs. MSA single (double)-mutant probabilities, verifying that the inferred fields and couplings for both MPF and final model with additional BML refinement, labeled MPF-BML, can accurately reproduce the single (double)-mutant probabilities. (*C*) Scatter plot of the connected correlations (covariances), demonstrating the benefits of BML refinement over MPF alone. (*D*) Probability of number of mutations, verifying that the inferred fields and couplings after BML, but not after MPF alone, accurately reproduce higher order statistics beyond the single- and double-mutant probabilities (which were not inputs to the inference procedure).

We note that the predicted combining factor $\phi = 0.95$, established at the initial variable combining phase, produced a landscape that achieved the strongest correlation with fitness values compared with other values of ϕ (*SI Appendix*, Table S3).

Inferred Couplings Predict Contact Residues. We expect that strongly interacting pairs of residues should be associated with strong couplings. For example, we previously showed in other HIV proteins that strong couplings are associated with compensatory mutations or those that interact synergistically to make the double mutant especially unfit (22, 24, 25). We used the couplings inferred for gp160 to predict residues that are likely to be in contact in the protein's native state, which can be determined from crystal structures. The rationale is that residues that are in close spatial proximity in the protein's native state should have a greater influence on each other, and this effect can be quantified by a measure dependent on the corresponding couplings (37, 38). Similar approaches were originally developed to predict residues in contact in the 3D protein structure on the basis of their inferred interactions, and have recently been extended to identify sets of interacting proteins on the basis of strong inferred couplings (39, 40). As our model is inferred from in vivo data, we anticipate that the model should capture contacts in the gp160 trimer that constitutes the functionally important HIV spike, not just contacts between residues in monomers of gp120 or gp41. To test this, we compared the top predicted contacts (a function of our model couplings) with known contacts based on a crystal structure of SOSIP (41), a synthetic mimic of the native gp160 trimer [Protein Data Bank (PDB) ID code 5D9Q]. The results (Fig. 4*B*) demonstrate that the couplings in our inferred model are predictive of protein contacts for the gp160 trimeric spike of HIV. We found similar results for nine other crystal structures (*SI Appendix*, Fig. S3).

Although our inferred couplings can predict contacts well, apparent false positives are still observed. To examine this further, we considered the top 20 residue pairs that have the largest average product correction–direct information (APC-DI) scores, in which 8 of these are not predicted to be in contact (*SI Appendix*, Table S4). With the exception of one residue pair, these residues are located in the V2 or V4 loop, or are CD4 contacts. For the residues in the V2 loop or those that are CD4 contacts, a possible reason for these residue pairs having a high APC-DI score is due to the conformational changes that occur when gp160 interacts with CD4 and other coreceptors, which is not captured in the protein structures. Indeed, conformational changes due to gp160–CD4 interaction have been well documented for the V2 loop (42). These conformational changes may favor or disfavor certain pairs of mutations. For the 389–417 residue pair in the V4 loop, it is not clear why this pair has a high APC-DI score, although we note that these residues are located near the base of the loop.

Observed Mutations in Footprints of CD4 Binding Site bnAbs Have Low Fitness Costs. As further validation of our landscape, we investigated whether its predictions explain why certain mutations that change the binding of bnAbs are observed in circulating virus strains, and others are not. Based on crystal structures, we first determined the residues on gp160 that were in contact with a number of CD4 binding-site bnAbs (listed in *SI Appendix*, *Text 8*). The CATNAP online tool (43) is a database of experimental IC-50 measurements between panels of bnAb–virus pairs. To determine whether mutations in gp160 residues that were in contact with the bnAbs affected binding, we used a feature in CATNAP that uses Fisher's exact test to identify viral amino acids that are statistically associated ($P < 0.05$) with either high or low IC-50 scores. This analysis allowed us to determine a set

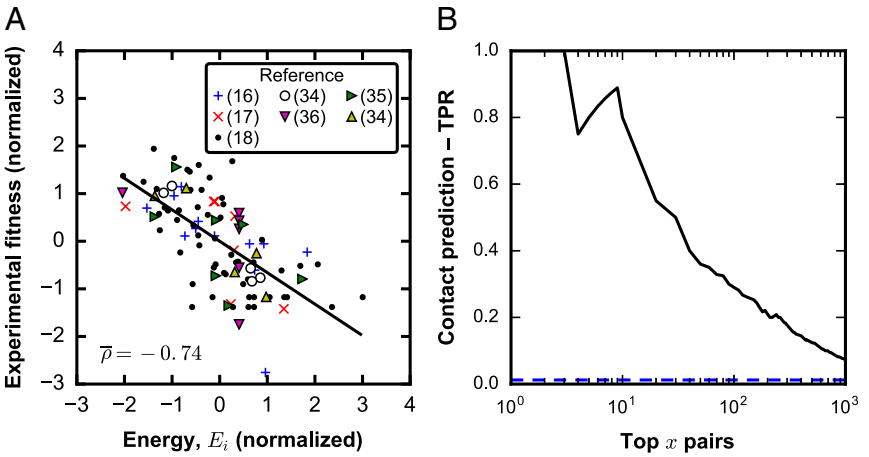

Fig. 4. (*A*) Normalized logarithm of fitness vs. normalized energy for all seven experimental datasets collected from the literature for a combining factor of $\phi = 0.95$. References for the datasets are shown in the legend. Normalization is performed by subtracting the mean of each dataset and dividing by the SD. A strong weighted correlation (*SI Appendix*, Eq. **S25**) is observed (see also *SI Appendix*, Fig. S1 for individual experiments). (*B*) Fraction of contacts in the top *x* predicted pairs which are actually in contact [true-positive rate (TPR)] vs. the top *x* predicted pairs as determined by a high value of the coupling constant (black line). True contacts are calculated from a crystal structure of the SOSIP trimer (5D9Q). These data show that the pairs with high predicted values of the coupling constants predict true contacts observed in the crustal structure of the SOSIP trimer. Also graphed is the total number of contacts (observed in the crystal structure) divided by the total number of pairs (horizontal blue line), which represents the probability of choosing a pair in contact purely by chance. Only pairs that are more than five residues apart in sequence space are considered, and out of these pairs, two residues are assumed to be in contact if they are <8 Å apart. PDB ID code 5D9Q.

of residues in gp160 contained in the footprints of CD4 binding-site antibodies that are statistically associated with a change in IC-50 for at least one mutant amino acid. These residues have at least one amino acid substitution pathway that changes the binding affinity to bnAbs. We analyzed only viral strains observed in the MSA that correspond to viable viruses.

At the identified set of residues, we further classify specific amino acid mutations as either "observed," defined as those mutations associated with a change in binding to bnAbs in the CATNAP data and observed in the MSA, or "unobserved," defined as mutations that are not in the CATNAP data (irrespective of whether they are observed in the MSA). We found that roughly 90% of these unobserved mutations are absent from the MSA (*SI Appendix*, Fig. S5*A*), and therefore are likely to be unviable. Furthermore, the largest frequency for an amino acid mutation present in the MSA but absent from CATNAP is small [i.e., less than 5% (*SI Appendix*, Fig. S5*B*)], while the amino acid mutations observed in the CATNAP data are representative of those observed in the MSA (*SI Appendix*, Fig. S5*C*). We expect, therefore, that observed mutations would be associated with lower fitness costs than unobserved mutations, since they are associated with viable viruses, and hence they are expected to escape the immune pressure with the least effect on the ability to propagate infection.

To assess whether the predictions of our fitness landscape are consistent with these expectations, we employed a measure that has been used before to predict the fitness cost of a mutation averaged over all sequence backgrounds (24). Specifically, we compute the change in energy associated with the mutation averaged over a Monte Carlo sample of diverse sequence backgrounds obtained using our fitness landscape to approximate:

$$\Delta E_{i,\mathbf{h}_f,\mathbf{J}_f}(a) = \sum_{\mathbf{x},x_i=0} \left[E_{\mathbf{h}_f,\mathbf{J}_f}(\mathbf{x}') - E_{\mathbf{h}_f,\mathbf{J}_f}(\mathbf{x})\right] p_{\mathbf{h}_f,\mathbf{J}_f}(\mathbf{x}), \qquad [5]$$

where $\mathbf{h}_f$ and $\mathbf{J}_f$ represent the final parameters after BML refinement, and the summation is over all sequences $\mathbf{x}$ having the consensus amino acid ("0") at the *i*th residue. Here, $\mathbf{x}'$ is identical to $\mathbf{x}$, except with the amino acid at residue *i* replaced by *a*. This measure quantifies the typical change in fitness upon introducing this mutation at the residue across various sequence backgrounds in circulating viruses. A positive fitness cost implies a decrease in fitness (22, 24). As predicted, the observed mutations at the residues that affect IC-50 are predicted to incur a lower average fitness cost compared with amino acid mutations that were not observed (Fig. 5*A*).

The difference in mean fitness cost between observed and unobserved mutations is due to the presence of some observed mutations with very low fitness costs, coupled with high fitness costs for many unobserved mutations. The existence of the low-fitness cost pathways that also likely abrogate binding implies that even canonically broad and potent bnAbs are imperfect. In patients who develop bnAbs upon natural infection, the viral quasispecies readily evolves escape mutations (1); escape mutations also eventually arise when bnAbs are administered passively (44), even when multiple bnAbs are used in concert (44, 45).

To further test our fitness landscape, we next focused on the gp160 residues in CD4 binding site antibody footprints that are not statistically associated with changes in IC-50 according to the analysis using CATNAP. We first examined whether the amino acid mutations at these residues that are observed are biochemically similar to each other and to the consensus amino acid at the residue. To determine this, for each residue, we calculated the biochemical similarity between the nonconsensus amino acid and the respective consensus amino acid, averaged over all sequences in the panels with nonconsensus amino acid at that residue (*SI Appendix*, Eq. **S25**). The similarity matrix (46) assigns to each pair of amino acids a similarity score ranging from 0 to 6, with matching amino acids assigned a value of either 5 or 6, where 6 is assigned to the more rare, unique amino acids (F, M, Y, H, C, W, R, and G) to reflect this. Mismatched amino acids are assigned values from 3 to 0 based on their polarity, hydrophobicity, shape, and charge (47). For each residue, the average biochemical similarity calculated above was compared with a simulated null model in which the nonconsensus amino acids observed in the bnAb–virus panels were randomly selected from all nonconsensus amino acids at that residue (*SI Appendix*, Fig. S5*D*). The average similarity was above the fifth percentile of the respective null model for 53% of the residues (*SI Appendix*, Fig. S5*D*), indicating that, on average, the amino acids sampled at these residues are more biochemically similar to the consensus amino acid than a uniformly selected random sample. The fact that the observed mutations are biochemically similar may explain why they did not significantly alter binding characteristics of the bnAbs. We then computed the fitness cost for the observed mutations at these residues and compared them with the mutations that were not observed as before. We expect that the latter mutations to biochemically different amino acids may have affected IC-50, but did not arise in viable viruses because of their high fitness costs; that is, the fitness cost associated with evolving escape mutations was prohibitive. Consistent with this expectation, our fitness landscape predicts that the fitness cost for the

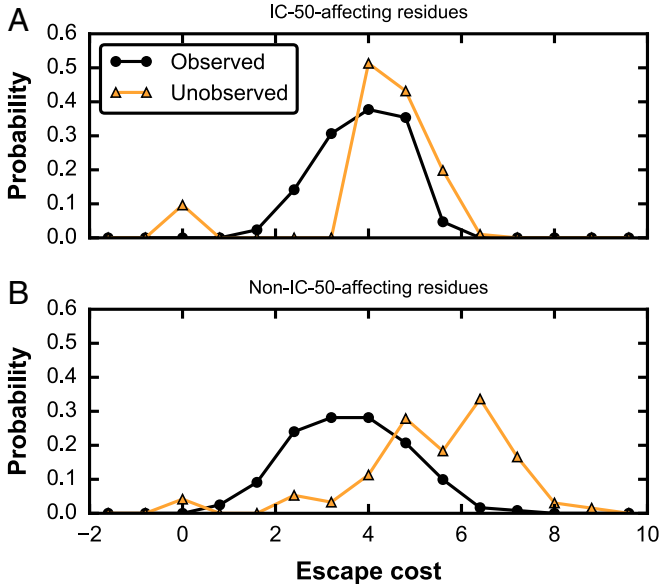

Fig. 5. Distribution of fitness costs for residues in the footprints of CD4bs-directed bnAbs. The footprint residues for a set of CD4bs-directed bnAbs (listed in *SI Appendix*, *Text 7*) were determined using crystal structures (as described in *SI Appendix*, *Text 7*). Next, the CATNAP online tool (43), a database of experimental IC-50 measurements between panels of bnAb–virus pairs, was used to identify amino acids within these footprints that are statistically associated with either high or low IC-50 for the bnAb–virus pairs tested. The footprint residues were separated into two sets of residues: the first set (*A*) consists of those residues for which at least one amino acid is statistically associated with a change in IC-50; the second set (*B*) consists of all other residues in the footprint where no such statistical associations with IC-50 were found. For both graphs, the fitness costs are calculated over all amino acids present in the CATNAP bnAb–virus panels (observed) and all amino acids not present in these CATNAP panels (unobserved). For the residues in *A*, the means of the observed and unobserved fitness cost distributions are 3.86 and 4.17, respectively ($P < 0.001$); for the residues in *B*, the means of the observed and unobserved fitness cost distributions are 3.61 and 5.42, respectively ($P < 0.001$); the *P* values were estimated by simulating from a null model in which the amino acids that were observed or unobserved are randomly permuted for each residue in the footprint.

observed mutations are lower than those for the unobserved mutations (Fig. 5*B*).

bnAbs That Target the CD4 Binding Site Contact Select Residues That Are Predicted to Be Associated with a High Fitness Cost upon Mutation. We computed the fitness cost associated with making mutations at all gp160 residues by carrying out Monte Carlo simulations to obtain an ensemble of sequences, all of which contained the consensus amino acid at a chosen residue. This ensemble of sequences was used to estimate the fitness cost for each particular mutation from consensus to a nonconsensus residue, as in Eq. **5** above. The results were then averaged over the mutant amino acids (24) to weight more likely mutations more highly as shown below:

$$\overline{\Delta E_i} = \frac{\sum_{a=1}^{\tilde{q}_i} \Delta E_{i,\mathbf{h}_f,\mathbf{J}_f}(a) \exp\left[-\Delta E_{i,\mathbf{h}_f,\mathbf{J}_f}(a)\right]}{\sum_{a=1}^{\tilde{q}_i} \exp\left[-\Delta E_{i,\mathbf{h}_f,\mathbf{J}_f}(a)\right]} \qquad [6]$$

where $\tilde{q}_i$ is the number of mutants at residue i after mutant combining (*SI Appendix*, *Text 2*). Eq. **6** provides us with an average fitness cost of evolving mutations at a particular residue, i, averaged over all sequence backgrounds in which the mutation might arise and over all mutant amino acids. We then superimposed this predicted residue-level fitness costs onto the SOSIP crystal structure. This can be visualized in a heat map depicting the fitness costs of residues that we determined to be accessible on the surface of SOSIP (Fig. 6*A*) (*SI Appendix*, *Text 6*). These results show that the map is very rugged, composed of a mixture of closely spaced low- and high-fitness cost residues. To obtain a clearer picture of whether low-fitness cost residues are predominant in any region of the size of a typical antibody footprint, we calculated the fitness cost (*SI Appendix*, Eq. **S24**) averaged over all surface residues within 12.5 Å of the chosen residue (the radius of a typical antibody footprint, as estimated from analyzing various bnAb–gp160 crystal structures). The corresponding heat map shows that fitness costs averaged over the antibody footprint are smooth over the entire surface on this scale (Fig. 6*B*). Ab-footprint–sized regions are typically dominated by low-fitness cost residues. Thus, most Ab responses that arise may be escaped via mutations. Even though the CD4 binding site has one of the largest fitness costs when averaged over a typical antibody footprint size (Fig. 6*B*), there still exist several residues associated with low fitness costs for mutations. Hence, if mutations at these residues lead to abrogation of Ab binding, then they present relatively easy escape pathways for the virus, consistent with the previous observation based on CATNAP data. We note that two other regions in Fig. 6*A* appear to have large numbers of high-escape cost residues: (*i*) the $\alpha 0$ and $\beta\bar{1}$ motifs of the C1 region of gp120, located in the trimer core near the intertrimer interfaces as well as the gp120–gp41 interface (48), and (*ii*) a region of gp41 containing portions of the H2 and membrane-proximal external region (MPER) motifs (49–52), although much of the MPER region is truncated in the 5D9Q crystal structure, including the epitopes for MPER-directed bnAbs 2F5 and 4E10 (51). However, the first region is sterically inaccessible to antibodies, while the second region contains a few highly variable residues that dominate the averaging in Eq. **6**, and thus on the scale of a footprint it is not very conserved.

We next examined the residues in the CD4 binding site region that are most strongly associated with the binding of the VRC01 class of antibodies. Li et al. (53) created an extensive set of alanine scanning mutants on a background of JRCSF, a CCR5-tropic clade B gp160 strain. For each alanine mutant, the authors measured binding affinity of full-length gp120 monomer to VRC01 relative to the wild-type virus. This assay led to the identification of residues where mutation to alanine led to significant decrease in VRC01–gp120 binding. The authors additionally conducted a neutralization assay in which whole mutant pseudoviruses are incubated with live target cells, and loss of infectivity is measured as a function of concentration of added CD4–Ig, a synthetic construct in which the CD4 domains 1 and 2 are fused to human IgG1 Fc domain. CD4–Ig competes with target cells to bind to the CD4 binding site of the pseudovirus, thus hindering infection of target cells. In this assay, if lower concentrations of CD4–Ig are required to inhibit infection, a decreased ability of the alanine mutant to infect via CD4-mediated interactions is indicated. In other words, a viral strain with a low CD4–Ig value has a low replicative fitness. The portion of these data corresponding to the gp160 residues in contact with VRC01 can be determined from crystal structures (*SI Appendix*, Table S5).

Of the residues on gp160 that were determined to be important for binding to VRC01, seven of these result in significant loss of binding (<33%) to gp120 upon mutation to alanine. While three of these seven (367, 368, and 457) are predicted by our landscape to have high fitness cost upon mutation, the remaining four (279, 371, 467, and 474) have much lower predicted fitness cost. This is seemingly at odds with the apparent high fitness cost of mutation to alanine for all seven residues, as measured experimentally (loss in sensitivity to CD4–Ig neutralization). This discrepancy can likely be explained by examining these residues at the amino acid level. For the four residues in question, the most common nonconsensus amino acid accounts for a large fraction of sequences in the MSA (*SI Appendix*, Fig.

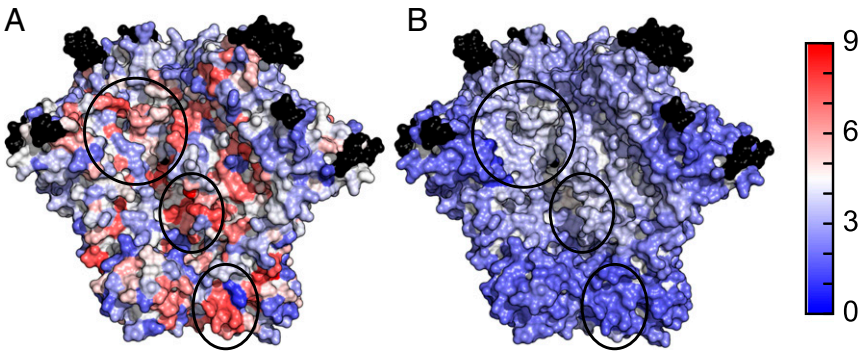

Fig. 6. Fitness costs superimposed on SOSIP Env gp160 trimer (PDB ID code 5D9Q). High-fitness cost residues (red) are more difficult for the virus to mutate, while low-fitness cost residues (blue) tend to be highly variable and easy to mutate. Residues in black are considered hypervariable and are not included in our model. (*A*) Fitness cost for each residue is averaged over all possible sequence backgrounds and amino acids at that residue. (*B*) Additionally averaged over surface residues within 12.5 Å of the residue. In both *A* and *B*, the regions circled in black, from *Top Left* to *Bottom Right*, correspond to the CD4bs, the $\alpha 0$ and $\beta\bar{1}$ motifs of the C1 region of gp120, and a region of gp41 containing portions of the H2 and MPER motifs, respectively.

S6), and it is not alanine. Thus, mutation to the most common nonconsensus amino acid is unlikely to result in a high fitness cost. Furthermore, in the CATNAP panel described in the previous section, neither the consensus nor the most common mutant amino acid is significantly associated with change in IC-50. Rather, the distributions of IC-50s for sequences containing either the consensus or the most common mutant amino acid are indistinguishable by two-tailed *t* test (*SI Appendix*, Fig. S6). This evidence suggests that mutation to the most common mutant amino acid may be insufficient to abrogate binding to VRC01. Our predicted fitness costs are likely artificially low, since Eq. **6** tends to weight low-fitness cost amino acid mutations highly during the averaging procedure and we infer that mutation to the most common nonconsensus amino acid does not incur a high fitness cost. If we instead calculate the fitness costs for these residues specifically to alanine, we predict much higher values (3.7, 4.2, 2.7, and 4.7, for residues 279, 371, 467, and 474, respectively). In contrast, for the three residues with high predicted fitness cost, the frequency of the most-common mutant amino acid was at most 0.2% in the MSA. Thus, although alanine was not the most-common mutant, the experimental fitness measurements made using alanine mutants were representative of typical escape costs at these residues. Most convincingly, 9 out of the top 10 residues with largest predicted fitness cost had an undetectable CD4–Ig value (*SI Appendix*, Table S5), further suggesting that our landscape provides an accurate measure of intrinsic fitness.

These results show that, consistent with observations regarding the VRCO1 class of antibodies, our fitness landscape predicts that bnAbs that target the CD4 binding-site region selectively bind to those residues associated with the largest fitness costs. They furthermore highlight the importance of describing fitness at the amino acid level of detail, not just residues—a fact that further highlights the importance of the fitness landscape that we have inferred.

Discussion

HIV is a highly mutable virus, and the design of both T cell and antibody arms of an effective vaccine would benefit from a knowledge of its mutational vulnerabilities as this would identify targets on the HIV proteome and viral spike where potent immune responses should be directed. Determining the fitness of the virus (ability to assemble, replicate, and propagate infection) as a function of the sequence of its proteins would help in this regard. Because of the significance of compensatory or deleterious coupling between mutations at multiple residues, in determining such a fitness landscape, it is important to account for the effects of coupling between mutations. Past work has been reasonably successful in determining such a fitness landscape for diverse HIV polyproteins by inferring the same from sequence data and testing predictions against in vitro experiments and clinical data (19, 22–25). This has not been true for the gp160 Envelope polyprotein that forms the virus's spike and is targeted by antibodies. This is due to the large diversity and size of gp160 (Fig. 1). We have reported a computational approach, based on the maximum-entropy method (19, 22), which includes a data-driven parameter reduction method, an initial estimation of the prevalence landscape parameters based on the principle of MPF, followed by a BML procedure to refine the parameters (Fig. 2). We applied this framework to infer the fitness landscape of gp160 from available sequence data. Predictions of the resulting model compare very well with published experimental data, including intrinsic fitness measurements, protein contacts in the SOSIP trimer, and known escape mutations that arise to evade antibodies that target the CD4 binding site.

A natural question, equally valid for any viral protein, is why one may see a strong relation between prevalence and fitness, and what are the fundamental processes underpinning this relationship. Motivated by this question, for internal HIV proteins that are targeted by cytotoxic T lymphocytes (CTLs), a mechanistic explanation has been proposed on the basis of simulation studies of HIV evolution and associated theory (22, 23, 26). Three key factors were identified for HIV proteins: (*i*) Because of the huge diversity of HLA genes in the population, very few regions of the HIV proteome are targeted by a significant fraction of humans. (*ii*) If a mutation is forced by the immune response of a particular individual and it incurs a fitness cost, then upon infection of a new patient who does not target the same region, during chronic infection the first mutation will revert. (*iii*) Unlike influenza, the HIV population has not been subjected to a few effective classes of natural or vaccine-induced immune responses by a significant fraction of humans across the world. Therefore, it has not evolved in narrowly directed ways to evade effective herd memory responses. (*iv*) Furthermore, the effects of phylogeny are ameliorated due to high rates of recombination (54). For these reasons, over reasonable mutational distances, HIV proteins are in a steady state and so amenable to inference of fitness landscapes using maximum-entropy approaches applied to sequence data. This is decidedly not true for the influenza population, which is driven far out of equilibrium (55, 56). The first of the above arguments, however, does not translate directly to gp160, which is primarily targeted by antibodies (57). While it is tempting to draw high-level analogies with the CTL-based phenomena on the basis that the diversity of antibodies generated against HIV is large, there are clear distinctions. For example, it is known that the V2 loop of

gp160 is estimated to be immunogenic in 20–40% of infected patients, while the V3 loop induces antibodies in essentially all infected patients (58). This would seem to promote more directed evolution compared with the internal proteins under CTL pressure. However, it is possible that different individuals evolve antibodies that target different residues in the V2 and V3 loops. It is also true that peptides derived from gp160 are targeted by CTLs, and that antibody responses are often directed toward debris from the fragile HIV spike. These reasons may underlie why we see excellent correspondence between our inferred model for gp160 and experimental measurements. Nonetheless, the mechanistic reasons remain unclear for gp160. Resolving these basic principles governing the strong prevalence–fitness correspondences identified in this paper for gp160 is worthy of future study.

We also considered an optimized model using the fields only, obtained by fitting only the single-mutant probabilities. We observe a modest gain in using the model with couplings, compared with the fields model; that is, the fields model has a correlation of $\overline{\rho} = -0.69$ compared with the correlation of $\overline{\rho} = -0.74$ using our computational framework.

The reason the comparison with in vitro fitness measurements is not much worse for the model with fields only is because the coupling matrix is relatively sparse, and so unless fitness measurements are carried out with only the subset of mutants with large couplings, the fields are dominant. No such preselection was done in the data on in vitro fitness that we compared our predictions against. The value of the coupling parameters is seen in several contexts. For example, protein contacts simply cannot be predicted without the couplings because the physical interactions between these residues induce a correlated mutation structure. More importantly, the fitness consequences of the couplings can be very significant for the evolution of HIV in vivo (the situation of ultimate interest), as described below.

Barton et al. (24) studied a cohort of HIV-infected individuals and used the inferred fitness landscapes of HIV proteins (other than gp120) to predict how long it took for the virus to escape from the initial T cell immune pressure in individual patients. When using single-residue entropy alone (i.e., fields only), there was only a 15% correlation between single-residue conservation and escape time. However, when they carried out dynamic simulations of the evolution using the fitness landscape that includes coupling parameters, this correlation was 72% and statistically significant. The reason for this dramatic increase in performance upon including the couplings is that evolutionary trajectories in vivo can sample many potential compensatory pathways that may aid escape due to the high rate of replication and mutation of HIV. Therefore, pathways characterized by particularly strong compensatory couplings can be accessed. Similarly, Barton et al. (24) showed that dramatic differences in escape time for patients targeting the same epitope could be explained by differences in the background mutations contained in the sequence of the virus strains that infected these patients. For example, if the background mutations had strong negative couplings with the ultimate escape mutation, the escape mutation took much longer to evolve and take over the population. Therefore, in the context of the real in vivo situation of interest, including the couplings is very significant. This point has also been emphasized in other contexts where compensatory mutations have been observed (16–18, 59, 60).

While our method was developed to address the specific challenges posed by the gp160 protein, the approach is general and may be applied to other high-dimensional maximum-entropy inference problems. For example, it may be directly applied to estimate the prevalence landscape of other HIV proteins, or proteins of other viruses. As an example, we applied the framework to p24, a relatively conserved internal protein of HIV. The landscape with ~80,000 parameters was inferred very quickly (~2.3 h for MPF and 1.2 h for the subsequent BML), and predictions from this landscape compared favorably with experimental fitness values (19), returning a strong Spearman correlation of −0.8 (*SI Appendix*, Fig. S2). For p24, along with other HIV proteins, similar correspondences between prevalence and fitness have been obtained through other maximum-entropy inference methods (19, 22–24).

We showed that the fitness landscape of the surface-accessible parts of the ENV trimer is very rugged where a few residues that incur a large fitness cost upon mutation (including the effects of couplings between mutations) are surrounded by residues with a low fitness cost upon mutation (Fig. 6). Even for the relatively "conserved" CD4 binding-site region, only mutations at some residues are predicted to incur large fitness costs upon mutations. This is consistent with the coarse-grained models of antibody–virus binding employed in models of affinity maturation with variant antigens by Wang et al. (61) and Shaffer et al. (62). In contrast, in a related study, Luo and Perelson (62, 63) assumed that variant antigens contain epitopes composed of only conserved or only variable residues. Our observation about the spatial distribution of escape costs on the surface of HIV envelope suggests that antibody footprints will in general contain both conserved and variable residues. It is likely that these conserved residues need to be the principal targets of effective antibody responses that are difficult for the virus to evade. For this reason, the availability of our fitness landscape of HIV's envelope proteins is expected to aid the rational design of immunogens that could potentially induce bnAbs upon vaccination (1–7), as well as guide the choice of combination of known bnAbs that would prevent escape in humans undergoing passive antibody therapy (10).

Specifically, our fitness landscape can be clinically useful in the future for the selection of combination bnAb therapy. Much like in the realm of antiretroviral drugs, the use of multiple different bnAbs has the potential to prevent or delay escape mutations, as the viruses in the host must evolve escape mutations in combinations of binding sites of the bnAbs administered. Two bnAbs with epitopes that have many detrimental couplings between them in all possible sequence backgrounds are likely to delay escape in diverse patients with different virus strains due to the added difficulty of introducing simultaneous escape mutations in both epitopes. Furthermore, bnAbs that target epitopes containing residues wherein escape mutations can be easily compensated by other mutations should be avoided. The importance of considering the fitness landscape in an analogous context of antiretroviral drugs has been previously shown for protease by Butler et al. (25).

An additional application that is clinically relevant is the selection of variant antigens (e.g., variants of SOSIP) for use in a candidate HIV vaccine designed to elicit bnAbs. The induction of bnAbs via vaccination will likely require immunization with multiple variant antigens that share some conserved residues, but whose variable regions contain distinct mutations compared with each other. The fitness landscape inferred in this work can provide insight into which residues within these antigens should be mutated within candidate immunogens such that the induced bnAbs are cross-reactive to diverse circulating viral strains with different sequences.

Materials and Methods

SI Appendix, Text includes detailed descriptions of the data preprocessing (*SI Appendix, Text 1*), computational framework (*SI Appendix, Text 2*), fitness verification (*SI Appendix, Text 3*), residue contact prediction (*SI Appendix, Text 4*), comparison with other methods (*SI Appendix, Text 5*), and application of the landscape for characterizing the fitness cost of escape mutations from bnAbs (*SI Appendix, Text 6–8*). The landscape, processed MSA, and implementation of the computational framework are freely available on the following website: https://github.com/raymondlouie/MPF-BML.

ACKNOWLEDGMENTS. This research was funded by the Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard (to K.J.K., J.P.B., and A.K.C.), the Hong Kong Research Grant Council General Research Fund with Project 16207915 (to R.H.Y.L. and M.R.M.), and a Harilela endowment (to M.R.M.).

1. Kwong PD, Mascola JR, Nabel GJ (2013) Broadly neutralizing antibodies and the search for an HIV-1 vaccine: The end of the beginning. *Nat Rev Immunol* 13:693–701.
2. Burton DR, et al. (2012) A blueprint for HIV vaccine discovery. *Cell Host Microbe* 12: 396–407.
3. Klein F, et al. (2013) Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* 153:126–138.
4. Mascola JR, Haynes BF (2013) HIV-1 neutralizing antibodies: Understanding nature's pathways. *Immunol Rev* 254:225–244.
5. Szabo FK, Hoffman GE (2014) Autoreactivity in HIV-1 broadly neutralizing antibodies: Implications for their function and induction by vaccination. *Curr Opin HIV AIDS* 9: 224–234.
6. Verkoczy L, et al. (2013) Induction of HIV-1 broad neutralizing antibodies in 2F5 knock-in mice: Selection against membrane proximal external region-associated autoreactivity limits T-dependent responses. *J Immunol* 191:2538–2550.
7. Zolla-Pazner S (2014) A critical question for HIV vaccine development: Which antibodies to induce? *Science* 345:167–168.
8. Hansen SG, et al. (2011) Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* 473:523–527.
9. Hansen SG, et al. (2009) Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nat Med* 15:293–299.
10. Lu CL, et al. (2016) Enhanced clearance of HIV-1-infected cells by broadly neutralizing antibodies against HIV-1 in vivo. *Science* 352:1001–1004.
11. Barouch DH, et al. (2013) Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* 503:224–228.
12. Klein F, et al. (2012) HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* 492:118–122.
13. Haddox HK, Dingens AS, Bloom JD (2017) Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathog* 12:e1006114.
14. Allen TM, et al. (2005) Selective escape from $CD8^+$ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *J Virol* 79:13239–13249.
15. Ferrari G, et al. (2011) Relationship between functional profile of HIV-1 specific CD8 T cells and epitope variability with the selection of escape mutants in acute HIV-1 infection. *PLoS Pathog* 7:e1001273.
16. Troyer RM, et al. (2009) Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog* 5:e1000365.
17. Liu Y, et al. (2013) A sensitive real-time PCR based assay to estimate the impact of amino acid substitutions on the competitive replication fitness of human immunodeficiency virus type 1 in cell culture. *J Virol Methods* 189:157–166.
18. da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE (2010) Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics* 185:293–303.
19. Mann JK, et al. (2014) The fitness landscape of HIV-1 gag: Advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput Biol* 10: e1003776.
20. Doud MB, Bloom JD (2016) Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* 8:155.
21. Wu NC, et al. (2015) Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet* 11:e1005310.
22. Ferguson AL, et al. (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38: 606–617.
23. Barton JP, Kardar M, Chakraborty AK (2015) Scaling laws describe memories of host-pathogen riposte in the HIV population. *Proc Natl Acad Sci USA* 112:1965–1970.
24. Barton JP, et al. (2016) Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat Commun* 7:11660.
25. Butler TC, Barton JP, Kardar M, Chakraborty AK (2016) Identification of drug resistance mutations in HIV from constraints on natural evolution. *Phys Rev E* 93: 022412.
26. Shekhar K, et al. (2013) Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys Rev E Stat Nonlin Soft Matter Phys* 88: 062705.
27. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M (2016) Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol* 33:268–280.
28. Mora T, Walczak AM, Bialek W, Callan CG, Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107:5405–5410.
29. Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys Rev Lett* 106:090601.
30. Barton JP, De Leonardis E, Coucke A, Cocco S (2016) ACE: Adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* 32:3089–3097.
31. Sohl-Dickstein J, Battaglino P, DeWeese MR (2011) Minimum probability flow learning. *Proceedings of the 28th International Conference on Machine Learning* (International Machine Learning Society, Bellevue, WA), pp 905–912.
32. Barton J, Cocco S (2013) Ising models for neural activity inferred via selective cluster expansion: Structural and coding properties. *J Stat Mech Theory Exp* 2013:P03002.
33. Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference on Neural Networks* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), pp 586–591.
34. Lobritz MA, Marozsan AJ, Troyer RM, Arts EJ (2007) Natural variation in the V3 crown of human immunodeficiency virus type 1 affects replicative fitness and entry inhibitor sensitivity. *J Virol* 81:8258–8269.
35. Anastassopoulou CG, et al. (2007) Escape of HIV-1 from a small molecule CCR5 inhibitor is not associated with a fitness loss. *PLoS Pathog* 3:e79.
36. Kassa A, et al. (2009) Identification of a human immunodeficiency virus type 1 envelope glycoprotein variant resistant to cold inactivation. *J Virol* 83:4476–4488.
37. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
38. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–E1301.
39. Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci USA* 113:12186–12191.
40. Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. *Proc Natl Acad Sci USA* 113:12180–12185.
41. Jardine JG, et al. (2016) Minimally mutated HIV-1 broadly neutralizing antibodies to guide reductionist vaccine design. *PLoS Pathog* 12:1–33.
42. Sullivan N, et al. (1998) CD4-induced conformational changes in the human immunodeficiency virus type 1 gp120 glycoprotein: Consequences for virus entry and neutralization. *J Virol* 72:4694–4703.
43. Yoon H, et al. (2015) CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res* 43:W213–W219.
44. Klein F, et al. (2012) Broad neutralization by a combination of antibodies recognizing the CD4 binding site and a new conformational epitope on the HIV-1 envelope protein. *J Exp Med* 209:1469–1479.
45. Shingai M, et al. (2013) Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* 503:277–280.
46. McLachlan AD (1972) Repeating sequences and gene duplication in proteins. *J Mol Biol* 64:417–437.
47. Sneath PHA (1966) Relations between chemical structure and biological activity in peptides. *J Theor Biol* 12:157–195.
48. Pancera M, et al. (2010) Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proc Natl Acad Sci USA* 107:1166–1171.
49. Huang J, et al. (2014) Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature* 515:138–142.
50. Sun ZYJ, et al. (2008) HIV-1 broadly neutralizing antibody extracts its epitope from a kinked gp41 ectodomain region on the viral membrane. *Immunity* 28:52–63.
51. Krebs SJ, et al. (2014) Multimeric scaffolds displaying the HIV-1 envelope MPER induce MPER-specific antibodies and cross-neutralizing antibodies when co-immunized with gp160 DNA. *PLoS One* 9:e113463.
52. Julien J-P, et al. (2013) Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* 342:1477–1483.
53. Li Y, et al. (2011) Mechanism of neutralization by the broadly neutralizing HIV-1 monoclonal antibody VRC01. *J Virol* 85:8954–8967.
54. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. *Elife* 3:1–18.
55. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc Natl Acad Sci USA* 113:E1701–E1709.
56. Luksza M, Lässig M (2014) A predictive fitness model for influenza. *Nature* 507:57–61.
57. Overbaugh J, Morris L (2012) The antibody response against HIV-1. *Cold Spring Harb Perspect Med* 2:a007039.
58. Zolla-Pazner S, Cardozo T (2010) Structure-function relationships of HIV-1 envelope sequence-variable regions refocus vaccine design. *Nat Rev Immunol* 10:527–535.
59. Brockman MA, et al. (2007) Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. *J Virol* 81:12608–12618.
60. Martinez-Picado J, et al. (2006) Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *J Virol* 80:3617–3623.
61. Wang S, et al. (2015) Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* 160:785–797.
62. Shaffer JS, Moore PL, Kardar M, Chakraborty AK (2016) Optimal immunization cocktails can promote induction of broadly neutralizing Abs against highly mutable pathogens. *Proc Natl Acad Sci USA* 113:E7039–E7048.
63. Luo S, Perelson AS (2015) Competitive exclusion by autologous antibodies can prevent broad HIV-1 antibodies from arising. *Proc Natl Acad Sci USA* 112:11654–11659.