Supporting Information

Barton et al. 10.1073/pnas.1415386112

SI Text

Sequence Data. Our data consist of a set of multiple sequence alignments (MSAs) of HIV-1 clade B sequences for the proteins Gag (2,398 sequences), Nef (1,948 sequences), protease (9,396 sequences), and integrase (2,488 sequences), obtained from the Los Alamos National Laboratory HIV sequence database (www.hiv.lanl.gov). For protease, we selected only sequences from drug-naïve patients to minimize the effects of selection for drug resistance. Insertions with respect to the HXB2 sequence, a standard clade B reference sequence, were removed. To control sequence quality, we excluded sequences labeled as "problematic" in the database and removed sequences with gaps and/or ambiguous amino acids at ≥5% of sites. We selected one sequence per patient for inclusion in the MSA to prevent multiple sequences obtained from the same individual from biasing the sequence distribution.

For our analysis, we converted the amino acid sequences in the MSA into a binary form, as described in ref. 1. To do this, we first determined the most frequently observed amino acid at each site. Each amino acid sequence in the MSA was then converted into a vector of binary variables, $z = \{z_1, z_2, \ldots, z_N\}$, $z_i \in \{0, 1\}$, where N is the total length of the protein sequence. For each MSA sequence, each of the z_i is set equal to 0 (1) if the amino acid at site i matches (does not match) the consensus amino acid at the same site.

We then computed the average frequency of mutations at each site and at each pair of sites in the MSA, given by the following:

$$p_i^* = \frac{1}{B}\sum_{k=1}^{B} z_i^{(k)}, \quad p_{ij}^* = \frac{1}{B}\sum_{k=1}^{B} z_i^{(k)} z_j^{(k)}. \qquad \text{[S1]}$$

Here, B is the total number of sequences in the MSA, and the index k is a label for each MSA sequence. These mutation frequencies, or correlations, characterize the distribution of sequences in the MSA. Additionally, we measured the probability of observing a sequence with n mutations, obtained by counting the fraction of sequences in the MSA exhibiting a certain number of mutations.

Maximum Entropy Model. We seek to infer a model that captures the variability of sequences present in the MSA, including the correlations given in Eq. **S1**. The least biased, or maximum entropy, probabilistic model capable of reproducing the correlations is the Ising model, described by Eq. **1**. In recent years, similar models have been used to study a wide range of complex systems, from the activity of networks of neurons (2, 3) to antibody sequences in zebrafish (3, 4).

Mathematically, we must determine the set of fields h_i and couplings J_{ij} so that the Ising model correlations:

$$p_i = \sum_z z_i \frac{\exp(-H(z))}{Q}, \quad p_{ij} = \sum_z z_i z_j \frac{\exp(-H(z))}{Q}, \qquad \text{[S2]}$$

match those from the MSA (Eq. **S1**). Here, the sum over z is a sum over all 2^N binary sequences of length N. This problem is referred to as the inverse Ising problem. Formally, the solution to the inverse Ising problem can be found by determining the fields and couplings that minimize the cross-entropy between the Ising model and the data (5):

$$S^*(h_i, J_{ij}) = \log Q - \sum_{i=1}^{N} h_i p_i^* - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij} p_{ij}^*. \qquad \text{[S3]}$$

This is equivalent to maximizing the likelihood function, which is proportional to $\exp[-BS^*(h_i, J_{ij})]$. Due to the presence of the partition function Q, direct minimization of Eq. **S3** is computationally intractable. To solve the inverse Ising problem, we thus use the selective cluster expansion algorithm, a fast approximate method based on iteratively finding the minimum of a regularized version the cross-entropy,

$$S^*(h_i, J_{ij}; \gamma) = \log Q - \sum_{i\in\Gamma} h_i p_i^* - \sum_{\{ij\}\in\Gamma} J_{ij} p_{ij}^* + \frac{\gamma}{2}\sum_{\{ij\}\in\Gamma} J_{ij}^2, \qquad \text{[S4]}$$

on small, strongly interacting subsets of the full system, denoted here by Γ (5, 6). The regularization strength γ is expected to be of the order of $1/B$, the number of sequences in the MSA (6). Given a particular value of the regularization strength, we follow the procedure outlined in ref. 6 to infer the Ising model parameters.

Model Selection and Validation. Although we expect that the ideal regularization strength γ should be of order $1/B$, the optimal value is not known exactly. Thus, we tested values of the regularization strength ranging from $2/B$ to $1/(2\,B)$, ensuring that regularization strengths for all proteins are similar in magnitude. Different values for the regularization strength γ may yield slightly different values for the fields h_i and couplings J_{ij}, all of which provide a good fit to the correlations measured from the MSA.

We quantify goodness of fit for the correlations through the error terms:

$$\epsilon_{p_1} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(p_i - p_i^*)^2}{(\delta p_i^*)^2}}, \quad \epsilon_{p_2} = \sqrt{\frac{2}{N(N-1)}\sum_{i<j}\frac{(p_{ij} - p_{ij}^*)^2}{(\delta p_{ij}^*)^2}}, \qquad \text{[S5]}$$

where

$$\delta p_i^* = \sqrt{\frac{p_i^*(1-p_i^*)}{B}}, \quad \delta p_{ij}^* = \sqrt{\frac{p_{ij}^*(1-p_{ij}^*)}{B}}, \qquad \text{[S6]}$$

the expected SD of the empirical correlations due to finite sampling (6). To avoid overfitting, we select models for which ε_{p1} and ε_{p2} are ∼1; that is, the error between the model correlations and the empirical correlations from the MSA is of roughly the same size on average as the expected fluctuation in the correlations due to finite sampling.

To choose the model that best captures the properties of the empirical sequence distribution, among those for which $\varepsilon_{p1}, \varepsilon_{p2} \sim 1$, we search for the model that best reproduces higher-order correlations, which are not directly constrained by the inverse Ising inference problem. In particular, we choose the model that gives the best fit to the distribution of the number of mutations observed in a sequence, as measured by the SE between the true (MSA) and inferred (model) distributions. In all cases, we obtain an Ising model that accurately reproduces both the single- and double-mutation frequencies (Fig. S1 *A* and *B*) and the distribution of the number of mutations in a sequence (Fig. S1*C*).

Statistical Error on the Inferred Fields and Couplings. Finite sampling induces errors in the inferred fields h_i and couplings J_{ij}. We denote the set of fields and couplings minimizing Eq. **S3** by h_i^*, J_{ij}^*. In the limit that $B \to \infty$, the likelihood function is tightly concentrated around h_i^*, J_{ij}^*; thus, the probability of some deviation in the inferred fields and couplings $\Delta = \{h_1 - h_1^*, \ldots, h_N - h_N^*, J_{12} - J_{12}^*, \ldots, J_{(N-1)N} - J_{(N-1)N}^*\}$ can be expanded as follows (6):

$$P(\Delta) \approx \frac{\sqrt{\det \chi}}{(2\pi B)^{N(N+1)/2}} \exp\left(-\frac{B}{2}\Delta^T \chi\, \Delta\right), \qquad \text{[S7]}$$

where χ is the $N(N+1)/2$-dimensional covariance matrix of the Ising model, i.e., the matrix of second derivatives of the entropy with respect to the fields and couplings. The expected fluctuations in the inferred h_i and J_{ij} are thus as follows:

$$\delta h_i = \sqrt{\frac{(\chi^{-1})_{i,i}}{B}}, \quad \delta J_{ij} = \sqrt{\frac{(\chi^{-1})_{ij,ij}}{B}}. \qquad \text{[S8]}$$

Due to the large size of the covariance matrix, these quantities are difficult to compute directly for the proteins studied here. However, we have tested the robustness of the inferred fitness peaks and their properties to using a reduced amount of data from the MSA to compute the correlations, thus indirectly affecting the inferred h_i and J_{ij}, as well as direct perturbation of these parameters. Details on these tests are presented in subsequent sections.

Zero-Temperature Monte Carlo Simulation. We begin with the set of MSA sequences, converted into a binary form as described above. To identify physically relevant basins of attraction (i.e., local maxima of the fitness landscape), we perform a zero-temperature Monte Carlo simulation to ascend (descend) the fitness (energy) landscape, using each sequence from the MSA as the starting point of one simulation. At each step of the simulation, we compute the change in fitness resulting from flipping the value of each z_i from consensus (0) to mutant (1), or vice versa. If the change in fitness is negative in all cases, then the current sequence $\mathbf{z}$ is a local fitness maximum; thus, the simulation is stopped. Otherwise, we flip the value of z_i at the site i that increases the fitness the most. This process continues until a local maximum of the fitness is reached, where it is no longer possible to increase the fitness of the sequence by flipping one of the binary variables in $\mathbf{z}$. Each local maximum of the fitness defines a fitness peak. We determine the occupancy of each peak by counting the number of MSA sequences that evolve to the same local maximum sequence under zero-temperature Monte Carlo simulation.

Absence of Multiple Fitness Peaks in Uncorrelated Sequence Data. Due to finite sampling, fluctuations in the observed frequency of mutations, given in Eq. **S1**, lead to uncertainty in the inferred couplings and fields, as described above. Even if the frequency of mutations at each site is independent, nonzero couplings can be inferred, which could potentially lead to the inference of multiple fitness peaks where only one truly exists.

To verify that the peaks we observe are not due purely to finite sampling noise, we generated an artificial set of data for each protein by shuffling the amino acids in each column of the MSA, so that the frequency of mutations at each site is unchanged, but correlations between mutations at different sites are purely random. We then fit an Ising model to the shuffled data, following the same procedure as before. As before, the inferred Ising models accurately reproduce the single- and double-mutation frequencies as well as the distribution of the number of mutations in a sequence. However, the number of fitness peaks changes dramatically. For the proteins protease and integrase, all sequences from the shuffled MSA lie on the same peak. For Gag, we find only two fitness peaks (one occupied by 2,397 sequences, the other by a single sequence), and for Nef, we find only three (occupied by 1,190, 420, and 338 sequences). Thus, the fitness peaks we infer from the full MSA data are not the result of sampling noise alone.

Fitness Peaks Obtained from Monte Carlo Sampling of the Model. Rather than using sequences from the MSA as starting points for the zero-temperature Monte Carlo procedure to derive fitness peaks, we could instead use an artificial collection of sequences obtained by sampling from the inferred fitness landscape (Eq. **1**). Here, we show that this alternative method leads to equivalent results.

First, we generated a random sample of binary sequences, equal to the total number of sequences in the MSA, according to the probability distribution given in Eq. **1**. We then evolved each of these sequences through zero-temperature Monte Carlo simulations as described above. For Gag, we obtained 385 fitness peaks (all enriched in HLA-associated mutations), and for Nef, we found 477 fitness peaks (all enriched in HLA-associated mutations). The distribution of sequences across these fitness peaks also follows a power law (maximum-likelihood exponent of 1.05 for Gag and 0.98 for Nef). The fitness peaks found in this way are the same as, or very similar to, the fitness peaks described in the main text, and the number of sequences that is found to lie on each of them is similar (Fig. S5 *A* and *B*).

We obtained analogous results for the weakly immunogenic proteins. For protease, we found 4 fitness peaks (all enriched in HLA-associated mutations), and for integrase, we found 25 (all enriched in HLA-associated mutations, except for one peak sequence that contains no mutations). The fitness peaks that contain a large number of sequences are identical for both methods (Fig. S5 *C* and *D*).

Sensitivity of Inferred Fitness Peaks and Their Properties to Limited Sequence Data and Perturbations of the Fields and Couplings. Here, we show that the properties of the fitness peaks that we observe are stable under variations due to finite sampling as well as direct perturbations of the inferred couplings and fields.

To evaluate the sensitivity of the inferred peak properties to finite sampling, we repeated our analysis using a randomly selected sample of 80% of the MSA sequences for each protein to build to fitness landscape model. For Gag and Nef, we found a set of 403 and 522 fitness peaks, respectively, all of which were enriched in HLA-associated mutations. The distribution of sequences across the observed fitness peaks follows the same power law scaling as previously observed (maximum-likelihood exponent of 1.17 for Gag and 1.09 for Nef). Because the fitness landscape inferred with this subset of the full set of MSA sequences has different h_i and J_{ij} parameters, the fitness peaks we find are similar but not always exactly identical to those observed previously. Hamming distances between the fitness peaks found from the subset model and the corresponding peaks obtained from the full model are typically far smaller than the distance between fitness peaks in the full model, i.e., the shifted fitness peaks remain distinguishable (Fig. S5 *E* and *F*). We note that the fitness peaks that contain the most sequences tend to change the least between the subset and full models, as expected (Spearman correlation between number of sequences on a fitness peak in the subset model and Hamming distance to the corresponding peak in the full model of $r = -0.51$, $P = 1.9 \times 10^{-28}$ for Gag, and $r = -0.52$, $P = 4.8 \times 10^{-38}$ for Nef). The number of sequences that lie on each fitness peak is also similar to that obtained previously.

Similar results also hold for the weakly immunogenic proteins protease and integrase (Fig. S5 *G* and *H*). We obtained 7 fitness peaks for protease (all enriched in HLA-associated mutations) and 14 for integrase (all but one sequence, which contains no

mutations, enriched in HLA-associated mutations). Here, the most populous fitness peaks are recovered exactly; only the fitness peaks containing <10 sequences in the subset model differ between the subset and full models.

We also repeated our analysis using a perturbed fitness landscape, adding a normally distributed random variable with mean zero and SD σ to each field h_i and couplings J_{ij}. We chose σ to be 10% of the size of the corresponding field or coupling, a perturbation large enough that the correlations (Eq. **S2**) change substantially ($\varepsilon_{p1} = 109$, 122, 136, and 53, $\varepsilon_{p2} = 24$, 46, 42, and 9, for Gag, Nef, protease, and integrase, respectively). However, despite the change in the model correlations, the overall structure of the fitness (energy) landscape is preserved. In this case, we find 569 fitness peaks for Gag and 502 fitness peaks for Nef, all enriched in HLA-associated mutations. The distribution of sequences across these fitness peaks again follows a power law with maximum-likelihood exponent of 1.22 for Gag and 1.05 for Nef. Similar to the previous analysis using a subset of the MSA sequences, the Hamming distance between peak sequences in the perturbed model and their corresponding peaks in the original model is small compared with the typical distance between peaks. For the weakly immunogenic proteins, we find 6 fitness peaks for protease (all enriched in HLA-associated mutations) and 17 for integrase (all enriched in HLA-associated mutations, except for one sequence that contains no mutations). As before, all of the most populous fitness peaks are recovered exactly for these proteins.

Maximum-Likelihood Estimation of Power Law Exponents. We estimate the exponent characterizing the distribution of the number of sequences lying on a fitness peak by the method of maximum likelihood. The number of sequences that lie on a peak is a discrete variable, so we assume that the distribution follows a discrete power law of the form:

$$P(n) = \frac{n^{-\beta}}{Q}, \quad Q = \sum_{n=1}^{n_{\max}} n^{-\beta}. \qquad \textbf{[S9]}$$

The log-likelihood of the data given a particular value of the exponent β and the maximum number of sequences $n_{\max}$ is as follows:

$$\ell = -\beta \sum_{i=1}^{b} \log n_i - b \log Q, \qquad \textbf{[S10]}$$

where n_i is the number of sequences that lie on the ith fitness peak, and b is the total number of peaks. The exponent β that maximizes the log-likelihood depends only weakly on the specified maximum number of sequences. Choosing $n_{\max} = \infty$, we obtain estimates of $\beta = 2.040$ for Gag and $\beta = 2.022$ for Nef. With $n_{\max} = B$, the number of sequences in the MSA, we find $\beta = 2.038$ for Gag and $\beta = 2.020$ for Nef. Note that the exponent in the rank-frequency plot (Fig. 1) is $\beta - 1$, rather than β. This is because the rank of a peak containing n_b sequences is proportional to the fraction of peaks that contain n_b or more sequences, i.e., the cumulative distribution:

$$\text{rank}(n_b) \propto \int_{n_b}^{\infty} dn\, P(n). \qquad \textbf{[S11]}$$

Definition of HLA-Associated Mutations. In ref. 7, an extensive list of HLA-associated polymorphisms were identified in the HIV-1 proteins Gag, Nef, and Pol, a polyprotein that contains both protease and integrase. Amino acids at a site that were significantly enriched in the presence of a specific HLA allele were referred to as "adapted" HLA-associated polymorphisms. Similarly, amino acids that were significantly depleted in the presence of a particular HLA allele were referred to as "nonadapted." To fit with our binary sequence model, we consider a mutation at a specific site to be HLA-associated if an amino acid variant labeled as adapted at that site differs from the consensus amino acid, or if an amino acid listed as nonadapted at that site is the same as the consensus amino acid.

We quantified the enrichment of HLA-associated mutations in peak sequences by the probability (P value) of obtaining at least the observed number of HLA-associated mutations in each peak under the assumption that the mutated sites in the peak sequence were selected by chance. Consider a protein of length N, where mutations at K out of the N total sites are HLA-associated. If n sites are mutated in a peak sequence, the probability of obtaining k or more HLA-associated mutations is as follows:

$$p(k|n) = \sum_{m=k}^{K} \frac{\binom{K}{m}\binom{N-K}{n-m}}{\binom{N}{n}}. \qquad \textbf{[S12]}$$

Quantifying Properties of Fitness Peaks. To explore properties of the peak sequences, we computed their overlap with other peak sequences, as well as the average couplings among mutated sites between and within peak sequences. For each peak sequence z^b, let us define a set $b = \{i | z_i^b = 1\}$, i.e., a list of the sites in the peak sequence that are mutated. We define the overlap between two peaks indexed by b_1 and b_2 to be as follows:

$$\text{overlap}(b_1, b_2) = \frac{\|b_1 \cap b_2\|}{\|b_1 \cup b_2\|}, \qquad \textbf{[S13]}$$

the ratio of the number of mutations sequences b_1 and b_2 share in common, divided by the total number of mutations in both sequences. The average coupling between mutated sites in these peak sequences is as follows:

$$\bar{J} = \frac{\sum_{i \in b_1, j \in b_2} J_{ij}}{\|b_1\|\|b_2\| - \|b_1 \cap b_2\|}. \qquad \textbf{[S14]}$$

Here, the denominator gives the number of pairs $\{i, j\}$ that can be formed from $i \in b_1$ and $j \in b_2$ satisfying $i \neq j$. The average coupling between primary mutations (defined as sites which are mutated in less than 20% of the peak sequences) and/or secondary mutations (sites that are mutated in more than 20% of the peak sequences) mutated in a peak sequence can be computed similarly. We note that the specific threshold of 20% for classifying mutations as primary versus secondary is not important; small changes in the definition do not qualitatively affect the results. Measures of the overlap between fitness peaks and their average couplings are reported in Fig. 2 for the highly immunogenic proteins Gag and Nef, and in Fig. S4 for the weakly immunogenic proteins protease and integrase.

Assessing the Predicted Fitness Effects of Mutations. We examined how different types of mutations contribute to the energy (fitness) of sequences in the MSA. We expect that sets of mutations belonging to the same peak sequence are part of the same compensatory pathways; thus, they should collectively decrease the energy (increase the fitness) of sequences bearing these mutations more than would be expected for arbitrary mutations. To study this, we computed the change in energy of each MSA sequence z,

$$\Delta H_i = H(z_i) - H(z), \quad z_i = \{z_1, z_2, \ldots, z_{i-1}, 1 - z_i, z_{i+1}, \ldots, z_N\}, \qquad \textbf{[S15]}$$

obtained when each site i in a sequence is converted from consensus (0) to mutant (1), or vice versa. The values of ΔH_i obtained for each MSA sequence were then categorized into one of

four groups, based on whether the site i was mutated or not in the MSA sequence and in the corresponding peak on which that sequence lies:

- Site is mutated in the MSA sequence and in the corresponding peak. These values of ΔH_i quantify how the energy of the MSA sequence is affected by mutations in the corresponding peak sequence. As noted above, we expect such mutations lower the energy (increase the fitness).
- Site is not mutated in the MSA sequence but is mutated in the corresponding peak. These values quantify how the energy of the MSA sequence would change if additional mutations present in the corresponding peak sequence were added.
- Site is mutated in the MSA sequence, but not in the corresponding peak. These values quantify how the energy of the MSA sequence is affected by mutations outside the corresponding peak sequence. Although these mutations are not present in the peak sequence, they are presumably not arbitrary mutations and thus they should not have high energy (fitness) costs.
- Site is not mutated in the MSA sequence or in the corresponding peak. These values quantify the energy cost of adding "random" or "arbitrary" mutations, which is presumably high.

After categorizing each of the ΔH_i into one of these groups, we computed their average for each MSA sequence and compared the averages across sequences (Fig. S3).

We found that indeed mutations that are present in both an MSA sequence and in the corresponding peak sequence tend to lower the energy (increase the fitness) of the sequence more than the other categories of mutations. As expected, mutations that are present in either the MSA sequence or the peak sequence, but not both, also have typically low energy costs. The typical cost of arbitrary mutations is very high.

Idealized Model of Multiple Fitness Peaks. Here, we describe a simple, exactly solvable model that gives a phenomenological description of the properties of fitness peaks we have observed. We consider a collection of N binary sites or spins $z = \{z_1, z_2, \ldots, z_N\}$, with $z_i \in \{0, 1\}$. A set of s peaks labeled by $\alpha = 1, 2, \ldots, s$, are encoded in nonoverlapping subsets of N_α sites $\{i_1^\alpha, i_2^\alpha, \ldots, i_{N_\alpha}^\alpha\}$. The sites in each peak are constrained such that sequences belonging to the peak must have a certain fraction m_α of sites mutated:

$$\sum_{k=1}^{N_\alpha} z_{i_k^\alpha} = M_\alpha = m_\alpha N_\alpha. \qquad \textbf{[S16]}$$

In the limit of large N that we consider below, this is equivalent to the less strict condition that the average fraction of mutated sites is equal to m_α. This constraint makes no distinction between different sites in the same peak—any subset of $m_\alpha N_\alpha$ sites can be mutated with equal probability. In sequences that do not belong to the peak α, all N_α sites are equal to consensus (0). We assume that sequences belong to only one fitness peak.

In this model, the one- and two-point correlations are given by the following:

$$p_{i^\alpha} = f_\alpha m_\alpha, \quad p_{i_1^\alpha i_2^\alpha} = f_\alpha m_\alpha^2, \quad p_{i^\alpha i^\beta} = 0, \qquad \textbf{[S17]}$$

where f_α is the frequency with which sequences lie on peak α. These correlations can be enforced through a Hamiltonian of the same form as that given in Eq. **1**, with fields h_α for all sites in the same peak and couplings J_α between them. Additionally, we require large negative couplings $J_{\alpha\beta}$ between sites in different peaks to enforce the condition that sequences belong to just one peak. The partition function Q can then be expressed as a sum over independent contributions from each peak:

$$Q = \sum_{\alpha=1}^{s} Q_\alpha, \quad Q_\alpha = \sum_{\{z_{i^\alpha}\}} e^{-\beta H_\alpha}, \qquad \textbf{[S18]}$$

where the peak-specific Hamiltonian H_α is given by the following:

$$H_\alpha = -h_\alpha \sum_k z_{i_k^\alpha} - J_\alpha \sum_{k<l} z_{i_k^\alpha} z_{i_l^\alpha}, \qquad \textbf{[S19]}$$

and β is the inverse temperature.

In the limit of large N_α, the partition function will generically be dominated by just the largest term. That is, only sequences from the peak with the highest free fitness would be observed. In order for sequences from all peaks to be observed with finite probability at $\beta = 1$, we need the following:

$$\lim_{N_\alpha \gg 1} \log(Q_\alpha) = S_\alpha - E_\alpha = C, \qquad \textbf{[S20]}$$

for some constant C. Because the entropy S_α and energy E_α are extensive, $\log(Q_\alpha) \propto N_\alpha$, in order for peaks with different N_α (assumed to be independent variables) to contribute, the constant of proportionality C must be equal to zero to leading order in N_α. We note that, if the condition $\log(Q_\alpha) = 0$ held exactly for all peaks α, then the probability of observing a sequence lying on any fitness peak α would be the same, in contrast with the observed power law distribution of sequences across peaks. However, the relatively probability of observing sequences on different fitness peaks could be tuned by allowing $\log(Q_\alpha)$ to vary by small additive factors (subleading in N_α). Here, we assume for simplicity that Eq. **S20** holds with $C = 0$ for all peaks.

Because all configurations with the same number of mutations in a peak have the same energy, we have the following:

$$\begin{aligned} S_\alpha &= -N_\alpha (m_\alpha \log m_\alpha + (1 - m_\alpha)\log(1 - m_\alpha)), \\ E_\alpha &= -N_\alpha (h_\alpha m_\alpha + 2 N_\alpha J_\alpha m_\alpha^2), \end{aligned} \qquad \textbf{[S21]}$$

in the limit of large N_α. Equating S_α and E_α thus leads to Eq. **3**. In the large N_α limit the self-consistency condition $\langle z_{i^\alpha} \rangle = m_\alpha$ at $\beta = 1$ further requires the following:

$$-h_\alpha - 2 N_\alpha J_\alpha m_\alpha - \log m_\alpha + \log(1 - m_\alpha) = 0. \qquad \textbf{[S22]}$$

Combining the constraints in Eqs. **S21** and **S22**, we can derive formulas for the fields and couplings in terms of the fraction of mutated sites in each peak:

$$\begin{aligned} h_\alpha &= \log\left(\frac{m_\alpha}{1 - m_\alpha}\right) + \frac{2}{m_\alpha}\log(1 - m_\alpha), \\ J_\alpha &= -\frac{1}{N_\alpha m_\alpha^2}\log(1 - m_\alpha). \end{aligned} \qquad \textbf{[S23]}$$

Model of Viral Growth. To compute the time necessary for the mutations characterizing a fitness peak to emerge, we assume a starting condition where no sites in the sequence are mutated. We consider only the dynamics of the sites present in the peak sequence of interest; the rest of the system is ignored. Each site that is currently equal to consensus mutates with rate 1, and each site that is currently mutated reverts to consensus with rate $\alpha e^{-J(k-1)}$, where k is the current number of mutated sites. The exponential factor $e^{-J(k-1)}$ in the reversion rate quantifies the synergistic effect between mutations within the same peak sequence due to the positive couplings between them. For simplicity, we assume that couplings between all sites in the peak sequence are equal to the average coupling, denoted by J. The parameter α in the reversion rate quantifies the relative ease of making new mutations versus reverting existing mutations. Because these mutations are driven by immune pressure, we expect $\alpha < 1$.

With this choice of dynamics, the total rates of the accumulation of mutations p_k and the reversion of mutations q_k are as follows:

$$p_k = N - k, \quad q_k = k\ \alpha\ e^{-J(k-1)}. \qquad \textbf{[S24]}$$

Following ref. 8, one then obtains for each peak i the expected time T_i necessary to proceed from the initial configuration, where no sites are mutated, to the full set of N_i mutations, given in Eq. **5** in the main text. Using the total number of mutations and the average coupling between mutations for each fitness peak in the highly immunogenic proteins Gag and Nef, we computed a set of n_i, proportional to the number of expected sequences on each peak i, through Eq. **4**. We chose the parameter α such that the rank correlation between the measured number of sequences on each peak i and the computed number of sequences n_i was maximized, obtaining $\alpha = 0.133$ for Nef and $\alpha = 0.058$ for Gag. For both Gag and Nef, the correlation between the real and computed number of sequences on each peak depends only weakly on the value of α chosen in this range; thus, we use $\alpha = 0.1$ for both in Fig. 3*C*, where we set $f = 3.2$ for Gag and $f = 3.6$ for Nef. With this value of α, the Spearman rank correlation between the true and model prevalence is $r = 0.42$, $P = 1.4 \times 10^{-18}$ for Gag and $r = 0.40$, $P = 3.2 \times 10^{-20}$ for Nef (Fig. S8).

We also performed an identical calculation in the case that the number of mutations required was not N_i but some fraction $m_i N_i$, equal to the average number of mutations in the peak sequences shared by MSA sequences that lie on the same peak. Again, we obtain a power law scaling relation for similar values of the parameters, but over a limited range as described in the main text. The results are shown in Fig. S9*A*, where we have used the same value of α as in the previous case, $f = 4.9$ for Gag and $f = 6.9$ for Nef.

Incorporating variable mutation rates, to mimic that different types of amino acid mutations occur with different probabilities, can extend the range over which the power law holds. To do this, we assume that mutation rates are not identically equal to 1 as above, but rather chosen uniformly from the range [0.5, 1]. We can set the upper limit of the mutation rate to 1 without loss of generality, as this sets the overall timescale. The results that follow are insensitive to the precise value of the lower limit. Because the mutation rates are now unequal, we can no longer write a simple expression for the overall rate of accumulation and reversion of mutations as in Eq. **S24**, so we compute the T_i through simulation. Choosing the same $\alpha = 0.1$ and computing the average T_i for each fitness peak over 10^4 simulations, we obtain the set of n_i shown in Fig. S9*B*, using $f = 2.3$ for Gag and $f = 3.5$ for Nef.

1. Ferguson AL, et al. (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38(3): 606–617.
2. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007–1012.
3. Mora T, Bialek W (2011) Are biological systems poised at criticality? *J Stat Phys* 144(2): 268–302.
4. Mora T, Walczak AM, Bialek W, Callan CG, Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107(12):5405–5410.
5. Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys Rev Lett* 106(9):090601.
6. Barton J, Cocco S (2013) Ising models for neural activity inferred via selective cluster expansion: Structural and coding properties. *J Stat Mech* 2013(03):P03002.
7. Brumme ZL, et al. (2009) HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One* 4(8):e6687.
8. Murthy KPN, Kehr KW (1989) Mean first-passage time of random walks on a random lattice. *Phys Rev A* 40(4):2082–2087.

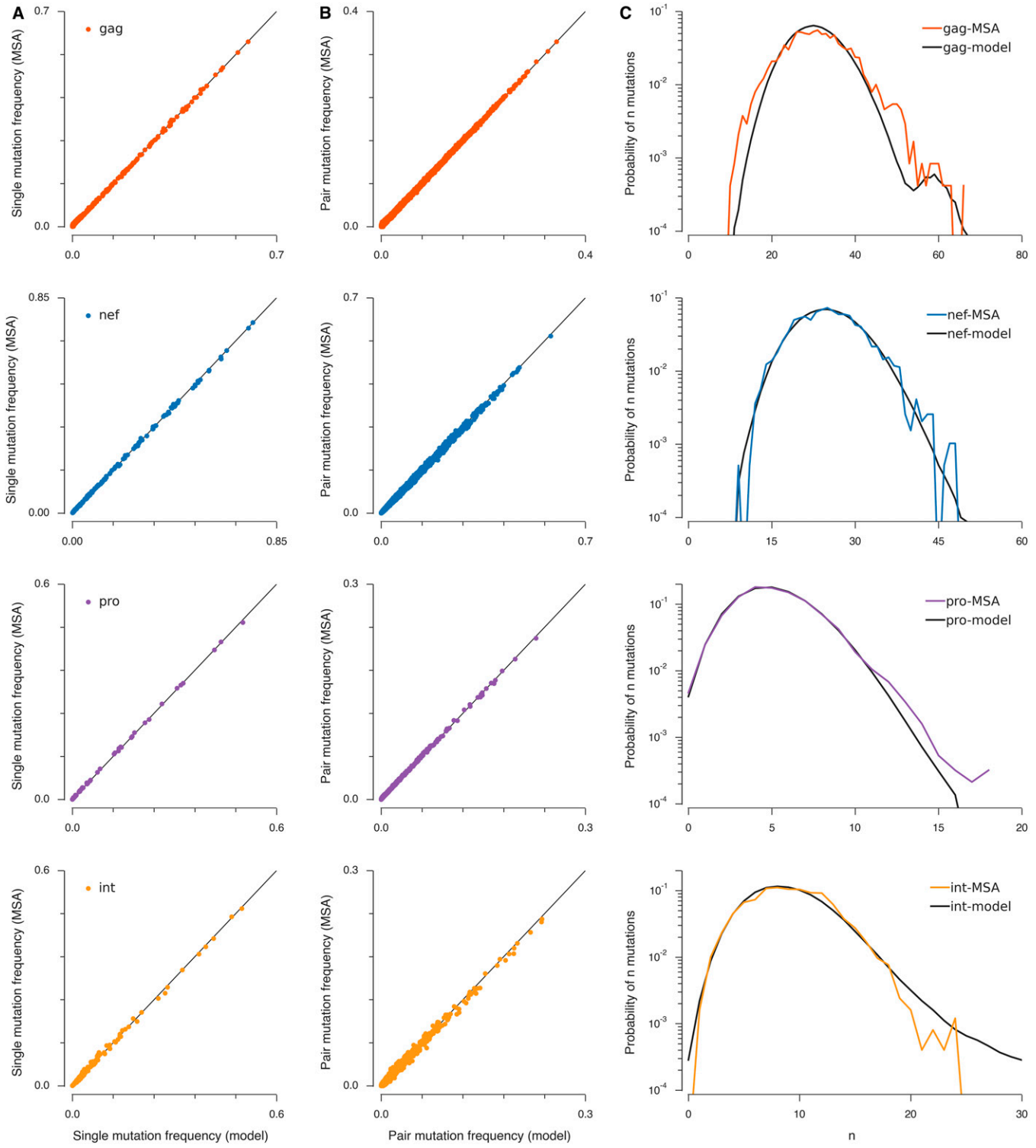

Fig. S1. The inferred Ising model accurately reproduces the frequency of single and double mutations as well as higher-order correlations in the sequence data. Comparison of single-mutation frequencies (*A*) and double-mutation frequencies (*B*) in the sequence data (Eq. **S1**) and the inferred Ising model (Eq. **S2**). (*C*) Comparison of the distribution of the number of mutations (relative to the consensus) observed in sequences between the sequence data, and the inferred Ising model. Note that the latter distribution is not directly constrained by the requirement that the single- and double-mutation probabilities in the inferred model match those in the data.

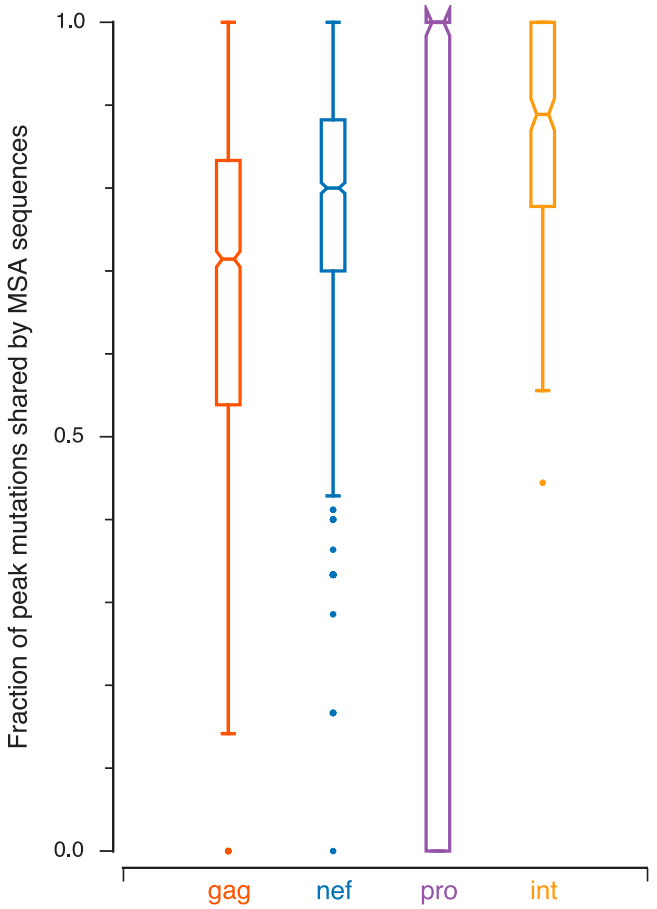

Fig. S2. A large fraction of the mutations that define a fitness peak are typically present in the sequences that lie on the same peak. Box plots show the median of the fraction of mutations in a peak sequence that are also present in the sequences from the MSA that belong to that fitness peak, with outliers plotted as dots. Means are 0.70 for Gag, 0.78 for Nef, 0.70 for protease, and 0.89 for integrase. Note that, because the peak that most protease sequences lie on has only one mutation, values of 0 or 1 are common.

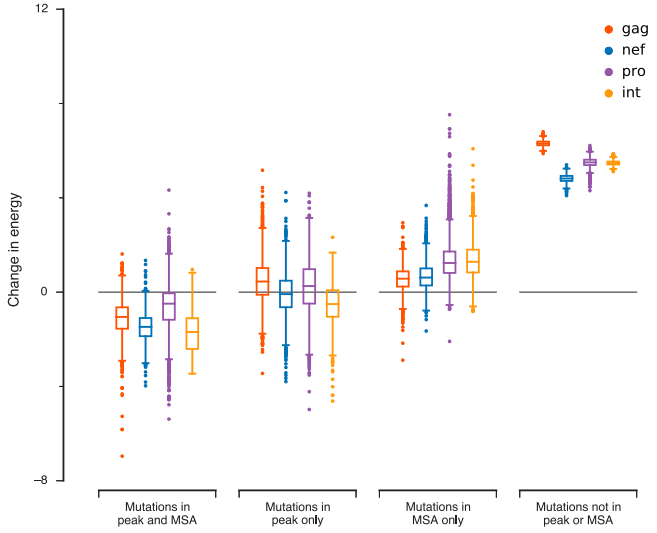

Fig. S3. Mutations vary in their effects on the energy of a sequence depending upon whether they are present in the MSA sequence, in the corresponding peak sequence, or both. Comparison of the contribution of point mutations to the energy of sequences from the MSA, categorized according to whether the particular mutation was present in the MSA sequence and its corresponding fitness peak (*SI Text*). Mutations present in both the MSA sequence and the sequence of the fitness peak on which it lies tend to decrease the energy (increase the fitness) of the sequence. Arbitrary mutations, present in neither the MSA sequence nor the corresponding peak sequence, tend to have very high energy costs. Mutations that are present in either the MSA sequence or the corresponding peak sequence, but not both, tend to have low energy costs.

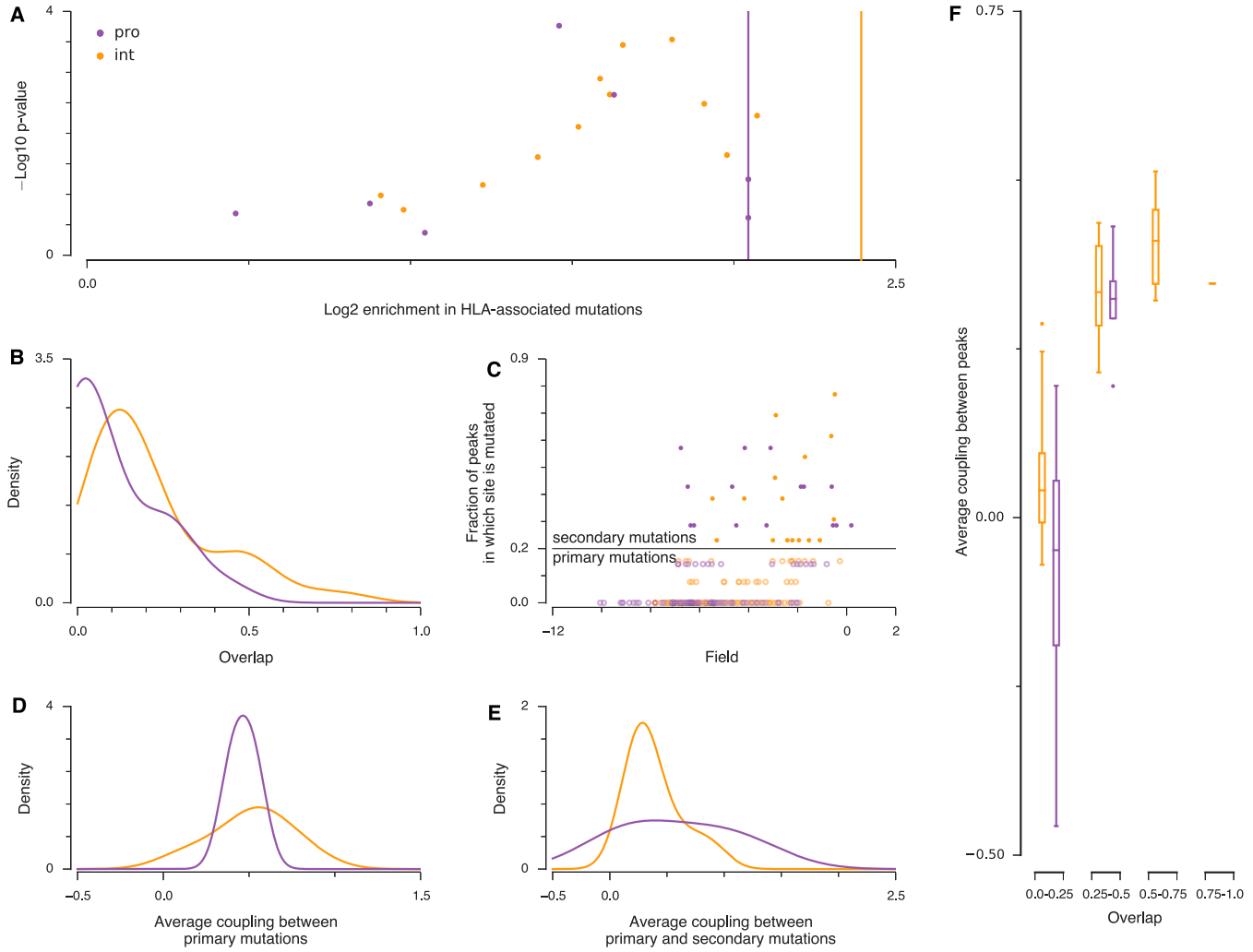

Fig. S4. Properties of fitness peaks in the weakly immunogenic proteins protease and integrase are similar to those obtained for Gag and Nef. (*A*) In protease and integrase, 7 of 7 (100%) and 12 of 13 (92%) peak sequences are enriched in HLA-associated mutations (definition in *SI Text*), respectively. Enrichment values are defined as the fraction of HLA-associated mutations in a peak sequence divided by the fraction of HLA-associated mutations in the whole protein. Vertical lines indicate maximum enrichment values, obtained if all mutations present in a peak sequence are HLA-associated. *P* values express the probability of obtaining at least as many HLA-associated mutations as actually observed in each peak sequence, assuming that these mutations were selected by chance (*SI Text*). (*B*) Most peaks are distinct, with little overlap between the sets of mutations present in other fitness peaks. Because very few peaks are observed for protease and integrase, small clusters of peaks that differ by containing slightly different sets of mutations cause a bump in the overlap distribution at large values of the overlap. These small clusters of peaks also appear in Gag and Nef. They are caused by negative interactions between some sets of mutations, which make them mutually incompatible. (*C*) Most mutations occur in a small fraction (20% or less) of peaks. A small fraction of mutations occur in many fitness peaks (15 for protease, 16 for integrase). (*D*) Average couplings between primary sites mutated within each peak sequence tend to be strongly positive. (*E*) Average couplings between primary and secondary sites mutated within each peak sequence are positive, but weaker than those between primary sites alone. (*F*) Average couplings between mutated sites in pairs of peaks are positive (compensatory) when the peaks strongly overlap, becoming more negative (deleterious) when the peaks are nearly disjoint.

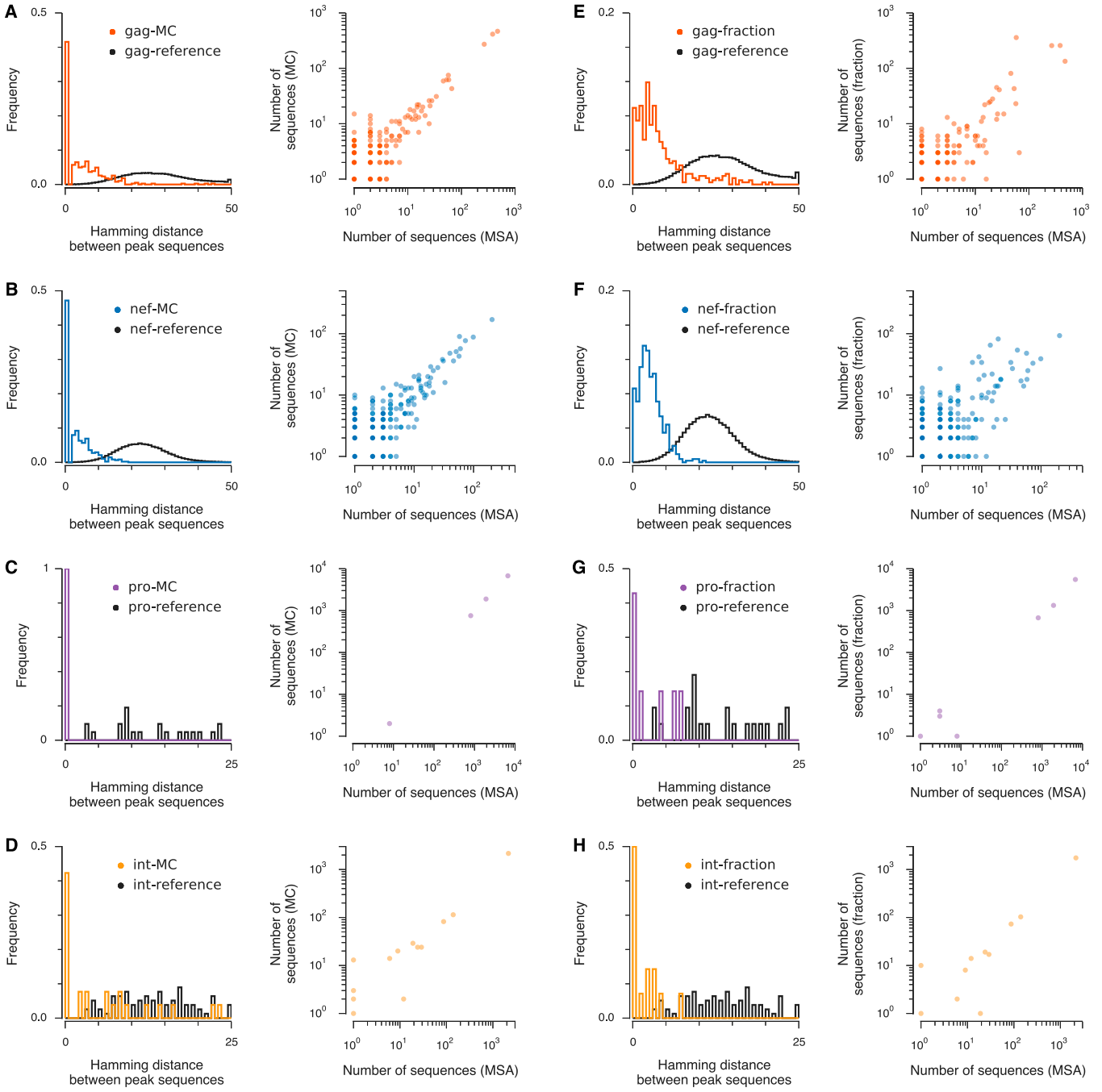

Fig. S5. Perturbed models yield similar fitness peaks. At *Left*, distribution of the Hamming distances between observed fitness peaks using binary sequences sampled from the inferred Ising model (Eq. **1**) and their corresponding fitness peaks found using the MSA sequences as starting points for the zero-temperature Monte Carlo procedure for the proteins Gag (*A*), Nef (*B*), protease (*C*), and integrase (*D*). For comparison, we show the distribution of Hamming distances between peak sequences in the original model. Fitness peaks found from a random sample from the inferred Ising model are the same, or very similar to, those found from the MSA sequences. At *Right*, comparison of the number of sequences that lie on each fitness peak found from the random sample of sequences and on the nearest corresponding fitness peak found from the MSA sequences. Differences between the distributions only emerge for fitness peaks that have a small number of sequences on them. The same plots are also shown for fitness peaks found using a model inferred from a fraction of the full set of data for Gag (*E*), Nef (*F*), protease (*G*), and integrase (*H*). The Hamming distance between fitness peaks found from the model using a fraction of the data and the original fitness peaks is small compared with the typical distance between peak sequences, indicating that fitness peaks remain distinct even when the model is inferred using limited data. The number of sequences that lie on each fitness peak is also consistent between the models inferred with a fraction of the data and those inferred using the full MSA.

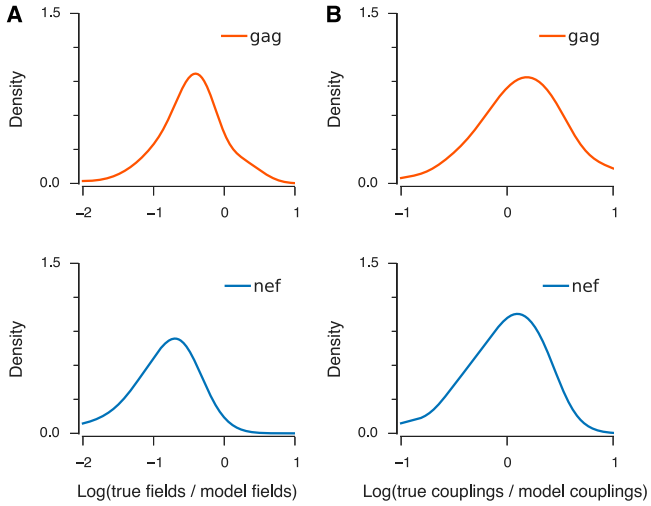

Fig. S6. Coupling parameters derived from a simple model of nonoverlapping fitness peaks are broadly consistent with those obtained from the maximum entropy model. (*A*) Kernel density estimate of the distribution of the log ratio of true average fields for each basin to those obtained from the basin model, with m_α taken to be the fraction of mutations in the basin sequence that are also present in the sequences from the MSA that fall into that basin (*SI Text*). The model overestimates the true magnitude of the fields. (*B*) The couplings derived from the model are similar in magnitude to the true couplings between sites in peak sequences.

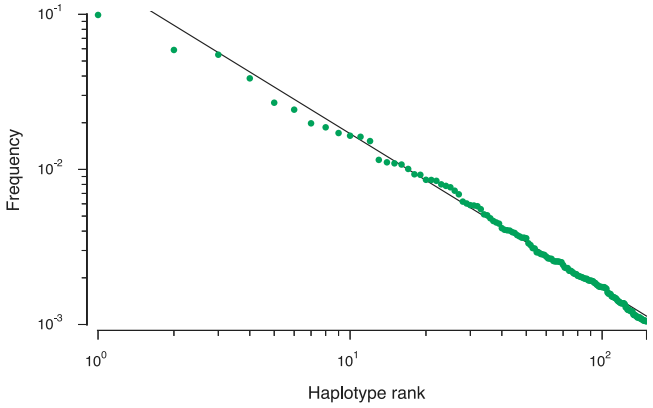

Fig. S7. The frequency of haplotypes in the human population also exhibits power law scaling. The frequency of the 150 most common haplotypes in the US population of European descent (1) is power law distributed with exponent ~1 (maximum-likelihood estimate, 0.97), similar to the observed distribution of HIV sequences across fitness peaks for the highly immunogenic proteins Gag and Nef (Fig. 1*A*).

1. Maiers M, Gragert L, Klitz W (2007) High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* 68(9):779–788.

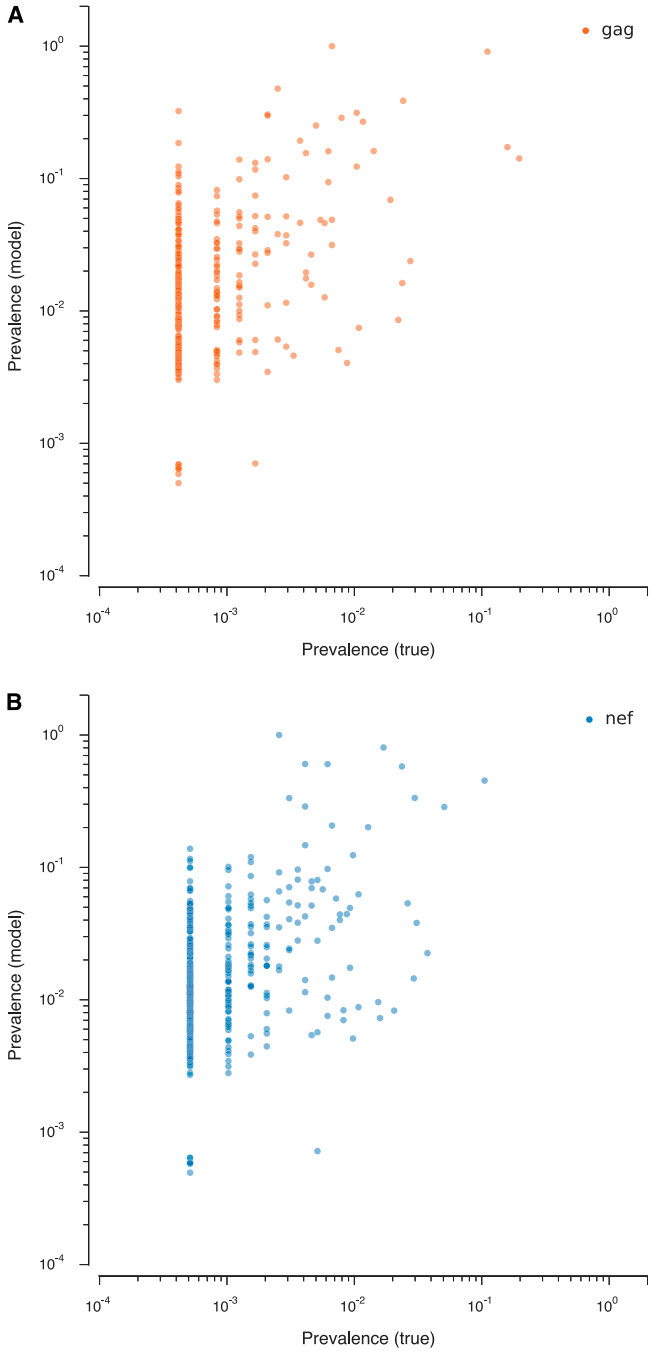

Fig. S8. The prevalence of sequences on each fitness peak in the viral growth model is correlated with the true number of sequences lying on the peak. Here, we show the true prevalence of sequences on each fitness peak versus the prevalence computed through the viral growth model (Eq. **4**) for Gag (*A*) and Nef (*B*). The computed prevalence is significantly correlated with the true prevalence (Spearman's $r = 0.42$, $P = 1.4 \times 10^{-18}$ for Gag, and $r = 0.40$, $P = 3.2 \times 10^{-20}$ for Nef).

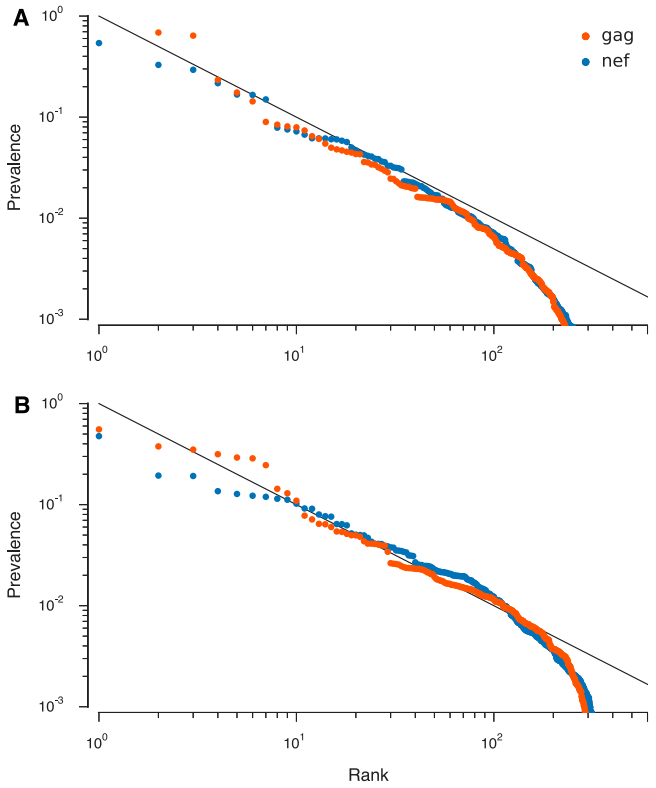

Fig. S9. Decreasing the number of mutations required for virus viability restricts the range of derived power law scaling, but range can be broadened through more variable dynamics. (*A*) Real sequences typically do not share all of the mutations in the peak sequence that they evolve to under zero-temperature Monte Carlo simulation (Fig. S2). Here, we show the derived distribution of sequences among basins using the simple model of viral growth (Eq. **4**), with the requirement that the number of mutations in each sequence needed to arise for the virus to be viable is equal to the average number of mutations in the peak sequence that are also present in the MSA sequences lying on that fitness peak. Because the width of the distribution of the entire set of peak sequence mutations is larger, the range of observed power law scaling is restricted. Normalization of the prevalence is arbitrary. For comparison, a power law with exponent 1 is shown in the background. (*B*) Incorporating more realistic variability in mutation rates broadens the range of the derived power law scaling (*SI Text*).

Table S1. Gag, protease, and integrase are similar at the level of single-site conservation

Protein	Mean conservation, %	Median conservation, %
Nef	87.7	96.4
Gag	93.8	99.0
Protease	94.7	99.4
Integrase	96.8	99.5

Conservation is defined for each site in an amino acid sequence as the fraction of sequences in the MSA where the amino acid at that site matches the consensus amino acid. Nef clearly displays increased variability compared with the other three proteins. Consideration of the full distribution of single-site conservation scores, beyond the mean and median shown here, yields similar results.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)