bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inference of epistatic effects and the development of drug resistance in HIV-1 protease

William F. Flynn[1,3], Allan Haldane[2,3], Bruce E. Torbett[4], and Ronald M. Levy*[2,3]

[1]Department of Physics and Astronomy, Rutgers University

[2]Department of Chemistry, Temple University

[3]Center for Biophysics and Computational Biology, Temple University

[4]Department of Molecular and Experimental Medicine, The Scripps Research Institute

July 1, 2016

Abstract

Understanding the complex mutation patterns that give rise to drug resistant viral strains provides a foundation for developing more effective treatment strategies for HIV/AIDS. Multiple sequence alignments of drug-experienced HIV-1 protease sequences contain networks of many pair correlations which can be used to build a (Potts) Hamiltonian model of these mutation patterns. Using this Hamiltonian model we translate HIV protease sequence covariation data into quantitative predictions for the stability and fitness of individual proteins containing therapy-associated

*Corresponding author: ronlevy@temple.edu

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutations which we compare to previously performed *in vitro* measurements of protein stability and viral infectivity. We show that the penalty for acquiring primary resistance mutations depends on the epistatic interactions with the sequence background and, although often destabilizing in a wildtype background, primary mutations are frequently stabilizing in the context of mutation patterns which arise in response to drug therapy. Anticipating epistatic effects is important for the design of future protease inhibitor therapies.

1 Introduction

The ability of HIV to rapidly mutate leads to antiretroviral therapy (ART) failure among infected patients. Enzymes coded by the *pol* gene play critical roles in viral maturation and have been key targets of several families of drugs used in combination therapies. The protease enzyme is responsible for the cleavage of the Gag and Gag-Pol polyproteins into functional constituent proteins and it has been estimated that resistance develops in as many as 50% of patients undergoing monotherapy [1] and as many as 30% of patients undergoing modern combination antiretroviral therapy (c-ART) [2].

The combined selective pressures of the human immune response and antiretroviral therapies greatly affect the evolution of targeted portions of the HIV-1 genome and give rise to patterns of correlated amino acid substitutions. As an enzyme responsible for the maturation of the virion, the mutational landscape of HIV protease is further constrained due to function, structure, thermodynamics, and kinetics [3–7]. As a consequence of these constraints, complex mutational patterns often arise in patients who have failed c-ART therapies containing protease inhibitors (PI), with mutations located both at critical residue positions in or near the protease active site and others distal from the active

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

site [7–10]. In particular, the selective pressure of PI therapy gives rise to patterns of strongly correlated mutations generally not observed in the absence of c-ART, and more therapy-associated mutations accumulate under PI therapy than under all other types of ART therapies [11–13]. In fact, the majority of drug-experienced subtype B protease sequences in the Stanford HIV Drug Resistance Database (HIVDB) have more than 4 PI-therapy-associated mutations (see Figure S1). Within the Stanford HIVDB are patterns of multiple resistance mutations, and in order to overcome the development of resistance, understanding these patterns is critical.

A mutation's impact on protein stability or fitness depends on the genetic background in which it is acquired. Geneticists call this phenomenon "epistasis". It is well understood that major drug resistance mutations in HIV protease destabilize the protease in some way, reducing protein stability or enzymatic activity, which can greatly alter the replicative and transmissive ability, or *fitness*, of that viral strain [4, 14]. To compensate for this fitness loss, protease accumulates accessory mutations which have been shown to restore stability or activity [8, 9, 15]. But it is unclear how the acquisition and impact of primary and accessory mutations are modulated in the presence of the many different genetic backgrounds observed, especially those present in the complex resistant genotypes that arise under inhibitor therapy.

Coevolutionary information derived from large collections of related protein sequences can be used to build models of protein structure and fitness [3, 16–20]. Given a multiple sequence alignment (MSA) of related protein sequences, a probabilistic model of the network of interacting protein residues can be inferred from the pair correlations encoded in the MSA. Recently, probabilistic models, called Potts models, have been used to assign scores to individual protein sequences which correlate with experimental measures

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of fitness [21–24]. These advances build upon previous and ongoing work in which Potts models have been used to extract information from sequence data regarding tertiary and quaternary structure of protein families [25–33] and sequence-specific quantitative predictions of viral protein stability and fitness [34, 35].

In this study, we show how such models can be constructed to capture the epistatic interactions involved in the evolution of drug resistance in HIV-1 protease. The acquisition of resistance mutations which accumulate under the selective pressure of inhibitor therapy leave many residual correlations observable in MSAs of drug-experienced sequences [12, 36, 37], and we use the pair correlations that can be extracted from MSAs to construct a Potts model of the mutational landscape of drug experienced HIV-1 protease. Due to the large number of model parameters and complex fitting procedure, we first provide several tests which demonstrate that our inferred model faithfully reproduces several key features of our original MSA including higher order correlations. We then compare the Potts model statistical energies with experimental measurements of fitness, including structural stability and relative infectivity of individual HIV protease variant sequences which contain resistance mutations. Finally, the Potts scores are used to describe the epistatic mutational landscape of three primary resistance mutations. We observe strong epistatic effects. The primary mutations are destabilizing in the context of the consensus (wildtype) background, but become stabilizing on average as the resistance mutations accumulate. Furthermore, the variance in the statistical energy cost of introducing a primary mutation also increases as resistance mutations accumulate; this heterogeneity is another manifestation of epistasis [38]. These findings provide a framework for exploring mutational resistance mechanisms using probabilistic models.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Results

2.1 Model inference and dataset

Given a multiple sequence alignment (MSA), we can infer a statistical model $P(\vec{\sigma})$ for the probability of finding a protein sequence with sequence identity $\vec{\sigma}$ which takes the form $P(\vec{\sigma}) \propto \exp(-E(\vec{\sigma}))$ from the statistical properties of the MSA. The maximum entropy model which reproduces the first and second order marginal distributions of the MSA, $P_i(\sigma_i)$ and $P_{ij}(\sigma_i, \sigma_j)$ of residue positions i and position pairs i, j, is given by the Potts Hamiltonian $E(\vec{\sigma}) = \sum_i h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j)$, where the fields $h_i(\sigma_i)$ and couplings $J_{ij}(\sigma_i, \sigma_j)$ represent the preference for residue σ_i at position i and residue pair $\sigma_i\sigma_j$ at positions i, j, respectively. The Potts model is fit to the bivariate marginals of the MSA such that it recovers the correlated pair information $C_{ij}(\sigma_i, \sigma_j) = P_{ij}(\sigma_i, \sigma_j) - P_i(\sigma_i)P_j(\sigma_j)$.

The Potts model captures epistatic effects; in contrast an independent model of a multiple sequence alignment can be constructed by summing the logarithm of the univariate marginals $E_{ind}(\vec{\sigma}) = \sum_i \log P_i(\sigma_i)$. As described in the following section, the ability to reproduce higher order marginals of the MSA (beyond second order) is a true predictive test of the Potts model, one which the independent model fails.

As described in the introduction, protease sequence evolution under protease inhibitor selective pressure produces correlations between amino acid substitutions that are larger in magnitude than those that occur in the absence of drug pressure (seen in Figure S2) [12, 37, 39]. Although correlations between drug-associated sites can be identified through analysis of drug-naive sequences [35, 36], the best model of the epistatic landscape of drug-resistance mutations comes from the correlations found in drug-experienced sequences.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Therefore, using a Stanford HIVDB MSA of 5610 HIV-1 subtype B drug-experienced protease sequences, we have inferred a Potts model using a Markov Chain Monte Carlo (MCMC) method implemented on GPUs (see Matierals and Methods and the supplemental information of Haldane et al. [32] for more details).

2.2 Recovery of the observed sequence statistics

We can gauge the accuracy of the model by examining how well the model reproduces various statistics of the MSA. The most direct test is the reproduction of higher order correlations observed in the multiple sequence alignment beyond pair correlations. Shown in Figure 1A is the recovery of the marginal probabilities of the most common subsequences observed in the dataset across varying subsequence lengths. The recovery of the bivariate marginals (pair frequencies) is not predictive but it demonstrates the quality of fit of the Potts model. The results shown in Figure 1 demonstrate that the Potts model is able to predict the frequencies of higher order marginals with accuracy. The Pearson correlation coefficient for the observed probabilities compared with the Potts model prediction remains above $R^2 \geq 0.95$ for subsequence lengths as large as 14. In contrast the independent model correlation coefficient is significantly worse ($R^2 \to 0.22$).

Figure 2 shows the probability distribution of sequences that differ from the consensus by k mutations as predicted by the Potts and independent models compared to the empirical distribution derived from the MSA. The Potts model predicts a distribution of mutations per sequence which is very close to the observed distribution whereas the independent model incorrectly predicts a multinomial distribution centered about 8 mutations from consensus. These predictions of various higher order sequence statistics provide strong evidence that the Potts model is not overfit.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Potts model also captures the observed statistics for larger subsequences, but as subsequence lengths increase, observed marginal probabilities in our MSA approach the sampling limit of the alignment ($1/N \approx 2 \times 10^{-4}$), meaning comparisons with the observed data at this level become dominated by noise. Tests with synthetic data (not shown) confirm that the discrepancy between observed higher order marginals and the Potts model are consistent with effects caused by the finite sample size (5610 sequences) of the MSA, and that the Potts model Hamiltonian is optimal given the data [32]. In the following section, we compare Potts model statistical energies with experimentally determined measurements of protease fitness.

2.3 Protease mutations, protein stability, and replicative capacity

Two experimental tests used to quantify the effects of protease mutations on viral fitness are thermal stability of the folded protein and replicative capacity [9, 40, 41]. Chang and Torbett [9] demonstrate that stability is compromised by the acquisition of primary mutations and this loss of stability can be rescued by known compensatory mutations, sometimes in excess of the reference stability. Muzammil et al. [40] and Louis et al. [41] have shown that patterns of up to 10 or more resistance mutations do not necessarily suffer from reduced fitness relative to the wildtype, and that non-active site mutations can lead to resistance in certain sequence contexts. In Figure 3A the change in statistical Potts energies, $\Delta E = E - E_{ref}$ corresponding to the sequences in these data sets is plotted versus the change in thermal stability and shows a strong correlation with the change in fitness as reflected by the change in melting temperature ($R = -0.85$, $p = 0.0003$). In contrast, the change in fitness computed using the independent model shows

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

no correlation (Figure S3A).

We have extracted results for viral replicative capacity in which 29 single mutants were studied by Henderson et al. [42] and an additional small set of more complex sequence variants [43] that were tested relative to the wildtype sequence. As with the stability measurements, we find the relative Potts energy correlates well with infectivity ($r = -0.64$, $p < 10^{-5}$), shown in Figure 3B. The same comparison using the independent model computed fitness again shows no predictive power (Figure S3B). Complementary to the RC assay presented in [42], Henderson et al. presented a SpIn assay and an additional assay measuring drug concentrations which inhibit protease function (EC50). Potts fitness predictions against these data are shown in Figure S4. While this additional comparison does not show statistically significant correlations, probably because the observed measurements span a much smaller range of values, they do exhibit the same negative trends observed in Figure 3. All data show in Figures 3, S3, S4 can be found in Supplementary Data 1.

The results presented here are reinforced by other recent studies of protein evolutionary landscapes [21–24] where varying measures of experimental fitness are compared to statistical energies derived from correlated Potts models constructed from multiple sequence alignments. The range of statistical energies and the correlation with fitness are qualitatively similar to those presented by Ferguson et al. [21] and Mann et al. [22] where statistical energies of engineered HIV-1 Gag variants generated using a similar inference technique are compared with replicative fitness assays. The same can be said for correlations between Potts scores and relative folding free energies of Beta Lactemase TEM-1 presented by Figliuzzi et al. [23]. While there are qualitative similarities between our results and related studies, differences exist in model inference procedures and, more im-

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

portantly, the evolutionary pressures that shaped these different mutational landscapes. Nonetheless, this collection of studies demonstrate that Potts model statistical energies correlate with the fitness of protein sequences in different contexts, including protein families evolving under weak selection pressure [23, 24], viral proteins evolving under immune pressure [21, 22], and now as presented here, viral proteins evolving under drug pressure.

2.4 Inference of Epistasis among therapy-associated mutations

The sequences present in the Stanford HIVDB have been deposited at many stages of HIV infection and treatment, showcasing a variety of resistance patterns spanning from wildtype to patterns of more than 15 mutations at PI-associated positions. In this section, we describe how Potts statistical energies can be used to infer epistatic effects on the major HIV protease resistance mutations.

Although all current PIs are competitive active site inhibitors, major resistance mutations can be found both inside and outside of the protease active site. V82 and I84 are positions inside the substrate cleft and major resistance mutations V82A and I84V have been shown to directly affect binding of inhibitors. L90 is a residue located outside of the substrate cleft and flap sites. Mutations at position 90, specifically L90M, have been shown to allow shifting of the aspartic acids of the active site catalytic triad (D25) on both chains, subsequently allowing for larger conformational changes at the dimer interface and active site cleft that reduce inhibitor binding [44–46].

Given a sequence containing one of the 3 mutants V82A, I84V, and L90M, we can determine the context-dependence of that mutation in its background by calculating the change in statistical energy associated with reversion of that mutation back to wildtype. This corresponds to computing $\Delta E = E_{obs} - E_{rev}$ where E_{obs} is the Potts energy of an

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

observed sequence with one of these primary mutations and E_{rev} is the Potts energy of that sequence with the primary mutation reverted to its consensus amino acid type. Due to the pairwise nature of the Potts Hamiltonian, this computation reveals a measure of epistasis for a sequence containing mutant $X \to Y$ at position k

$$\Delta E(\vec{\sigma}_{k,Y}) = h_k(Y) - h_k(X) + \frac{1}{2}\sum_{i \neq k}(J_{i,k}(\sigma_i, Y) - J_{i,k}(\sigma_i, X)) \quad (1)$$

where the pair terms $J_{i,k}$ are the couplings between the mutation site and all other positions in the background. When this measure is positive, the background imparts a fitness penalty for the reversion of the primary resistance mutation to the wildtype and when negative, the sequence regains fitness with reversion to wildtype. Using this measure, we compute ΔE for every sequence in our HIVDB MSA containing V82A, I84V, L90M and have arranged the energies versus sequence hamming distance from the consensus including only PI-associated sites, shown in Figure 4A,B,C respectively. We observe that as more mutations accumulate in the background, the fitness gain to revert the primary resistance mutation is lost and the primary mutation becomes stabilizing on average when enough mutations have accumulated. These crossover points are 6, 9, and 7 mutations for V82A, I84V, and L90M, respectively. Once a sufficient number of mutations have accumulated, the majority of sequence backgrounds are interconnected in such a way that the primary resistance mutation is entrenched, meaning a mutation to wildtype at that position is destabilizing, and the primary mutation becomes more entrenched as more background mutations are acquired. The effect is largest for L90M; for sequences containing a large number of PI-associated mutations, on average the L90M primary mutation is ≈ 100 times more likely than the wildtype leucine at position 90. In contrast, this primary mutation is ≈ 80 times less likely than the wildtype residue in the

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

subtype B consensus sequence background. The trend shared for V82A, I84V, and L90M is representative of the larger class of primary mutations; mutations such as V32I, M46L, I47V, G48V, I50V, I54V, L76V, and others become less destabilizing as the number of background mutations increases (see Figure S5).

As shown in Figure 4 the variance in ΔE, σ_E, grows as the hamming distance from the consensus initially increases. This is consistent with recent results suggesting that increasing or decreasing variance in fitness over time serves as a general indicator of an underlying epistatic mutational landscape [38]. That we observe this change in variance across hamming distance signals that the change of fitness when reverting primary resistance mutations to wildtype is due to the collective interaction of multiple residues.

Primary mutations are likely to revert to wildtype in the absence of inhibitors, We observed that reversion of a primary resistance mutation to wildtype is 10–100 times more probable when the total number of background mutations is small. Sequences near the crossover points are equally likely to revert the primary mutation as retain it. However for sequences with many mutations, retaining the primary resistance mutation can be 100–1000 times more favorable than reverting to wildtype. Increasing epistasis, as seen here as hamming distance from the wildtype grows, is associated with increased ruggedness of the fitness landscape [39, 47, 48]. Our results imply that this ruggedness may create local maxima on the fitness landscape which may become accessible as more resistance mutations accumulate in a sequence. These entrenched, highly resistant sequences with many mutations present a significant risk for the transmission of drug resistance to new hosts.

All together, this suggests that the acquisition of primary mutations relies on a complex network of interactions and that, while primary mutations are often deleterious to

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein fitness when acquired in a wildtype background, they become stabilizing genotypes in the presence of many PI-associated mutations. This also implies that primary mutations take on a context dependent accessory role, allowing for the acquisition of additional primary and accessory mutations.

3 Discussion

The evolution of viruses under drug selective pressure induces mutations which are correlated due to constraints on structural stability that contribute to fitness. The correlations induce epistatic effects, that is, the consequences of a particular mutation depends on the genetic background. Recently epistasis has become a focus for analysis in structural biology and genomics as researchers have begun to successfully link the coevolutionary information in collections of protein sequences with the structural and functional fitness of those proteins [19, 21–24]. In the studies presented herein, we used the correlated mutations encoded in a multiple sequence alignment of drug-experienced HIV-1 protease sequences to parametrize a Potts model of sequence statistical energies that can be used as an estimator of stability and relative replicative capacity of individual protease sequences containing drug resistance mutations. Using the statistical energy $E(\vec{\sigma})$ as a proxy for the fitness of sequence $\vec{\sigma}$, we find that the effects of primary resistance mutations vary significantly, depending on the background sequence in which they occur.

Understanding the epistatic relationships among drug resistance mutations in HIV has important implications for therapies. Viruses with many background mutations incur large penalties to revert primary resistance mutations to the starting genotype. Therefore, the reversion cost can entrench primary resistance mutations even in the absence of inhibitors. Our findings imply that highly mutated sequences could serve as reservoirs of

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

drug-resistance mutations after HIV transmission, which in turn would promote therapy failure in new hosts.

Recent publications have reported that mutations near or distal to Gag cleavage sites play a role in promoting cleavage by drug-resistant and enzymatically deficient proteases, by selecting for mutations that increase substrate contacts with the protease active site, altering the flexibility of the cleavage site vicinity, or by as of yet unknown mechanisms [8, 10, 49–52]. This suggests that viral coevolution of Gag with selective protease mutations may further stabilize multiple resistance mutations; thus, the analysis of protease mutation patterns can be extended to include amino acid substitutions within Gag and the Gag-Pol polyprotein. Furthermore, this type of analysis is not limited to protease and may be used to study the development of resistance in other HIV drug targets, such as reverse transcriptase and integrase, as well as other biological systems that develop resistance to antibiotic or antiviral therapies.

The Potts model is a powerful tool for interrogating protein fitness landscapes. The analysis presented here provides a tractable framework to examine the structural and functional fitness of individual viral proteins under drug selection pressure. Elucidating how patterns of viral mutations accumulate and understanding their epistatic effects has the potential to have an impact on the design and evaluation of the next generation of c-ART inhibitors and therapies.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4 Materials and Methods

Sequence Data

Sequence information (as well as patient and reference information) was collected from the Stanford University HIV Drug Resistance Database (`http://hivdb.stanford.edu`) [11] using the Genotype-Rx Protease Downloadable Dataset (`http://hivdb.stanford.edu/pages/geno-rx-datasets.html`) that was last updated on 29/04/2013 (there now exists a more recent sequence alignment updated May 2015). There are 65,628 protease isolates from 59,982 persons in this dataset. From this dataset, 5,824 drug-experienced, subtype B, non-mixture, non-recombinant, and unambiguous sequences were extracted. Sequences with more than 1 gap and MSA columns with more than 1% gaps (positions 1–5 and 99) were removed, resulting in $N = 5,610$ sequences of length $L = 93$.

For the comparison made in Figure S2, drug-naive subtype B non-mixture, non-recombinant, and unambiguous sequences were extracted from the same downloadable dataset. As with drug-experienced sequences, gap-containing sequences and columns were removed, resulting in 13,350 sequences of length 89.

Mutations considered PI-associated were extracted from [53]: L10I/F/V/C/R, V11I, G16E, K20R/M/I/T/V, L24I, D30N, V32I, L33I/F/V, E34Q, M36I/L/V, K43T, M46I/L, I47V/A, G48V, I50L/V, F53L/Y, I54V/L/A/M/T/S, Q58E, D60E, I62V, L63P, I64L/M/V, H69K/R, A71V/I/T/L, G73S/A/C/T, T74P, L76V, V77I, V82A/F/T/S/L/I, N83D, I84V, I85V, N88D/S, L89I/M/V, L90M, I93L/M.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Marginal Reweighting

Weights (w^k) reciprocal to the number of sequences contributed by each patient were computed and assigned to each sequence. With these weights, estimates of the bivariate marginal probabilities were computed from the MSA of N sequences:

$$P_{ij}(\sigma_i, \sigma_j) = \frac{1}{N} \sum_{k=1}^{N} w^k \delta\left(\sigma_i^k, \sigma_i\right) \delta\left(\sigma_j^k, \sigma_i\right) \tag{2}$$

where σ_i^k is the residue identity at position i of the kth sequence $\vec{\sigma}^k$, $0 < w^k \leq 1$ is the weight of sequence k, and delta $\delta(\alpha, \beta)$ equals one if $\alpha = \beta$ and is otherwise zero.

Otherwise, all sequences are assumed independent; no reweighting was done to account for shared ancestry among these sequences. Phylogenetic trees of drug-naive and drug-treated HIV-infected patients have been show to exhibit star-like phylogenies [39, 54], and thus phylogenetic corrections are not needed. Further, phylogenetic corrections based on pairwise sequence similarity cut-offs of 40% of sequence length or more as are common in studies utilizing direct coupling analysis (DCA) [25–27] of protein families would drastically reduce the number of effective sequences in our MSA and would lead to mischaracterization of the true underlying mutation landscape. Potts models of other HIV protein sequences under immune pressure have been parameterized with no phylogenetic corrections [21, 22].

Alphabet Reduction

It has been shown that "reduced alphabets" consisting of 8 or 10 groupings of amino acids capture most of the information contained in the full 20 letter alphabet [55]. We expand on this notion by computing an alphabet reduction that has the least effect on the

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

statistical properties of our MSA. In the context of model building, a reduced alphabet decreases the number of degrees of freedom to be modeled. This leads to a more efficient model inference [30, 32].

Given the empirical bivariate marginal distribution for each pair of positions in the MSA using 21 amino acid characters (20 + 1 gap), the procedure begins by selecting a random position i. All possible alphabet reductions from 21 to 20 amino acid characters at position i are enumerated for every pair of positions ij, where $j \neq i$, by summing the bivariate marginals corresponding to each of the 210 possible combinations of amino acid characters at position i. The reduction which minimizes the root square mean difference (RMSD) in mutual information (MI) content:

$$\sqrt{\frac{1}{N}\sum_{ij}\left(\mathrm{MI}_{ij}^{Q=21} - \mathrm{MI}_{ij}^{Q=Q'}\right)^2} \tag{3}$$

between all pairs of positions ij with the original alphabet size $Q = 21$ and reduced alphabet size $Q = 20$ is selected. The alphabet at each position i is reduced in this manner until all positions have position-specific alphabets of size $Q = 20$. This process is then repeated for each position by selecting the merger of characters which minimizes the RMSD in MI between all pairs of positions ij with the original alphabet size $Q = 21$ and reduced alphabet size $Q = Q'$, and is stopped once $Q = 2$.

Due to residue conservation at many loci in the HIV protease genome, the average number of characters per position is 2, and several previous studies of HIV have used a binary alphabet to extract meaningful information from sequences [10, 12, 21, 34]. However, using a binary alphabet marginalizes potentially informative distinctions between amino acids at certain positions, especially PI-associated sites, that acquire multiple mutations from the wildtype. We found that an alphabet of 4 letters substantially reduces the

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequence space to be explored during the model inference while providing the necessary discrimination between different types of mutant residues at each position. Additionally, the information lost in this reduction is minimal; Pearson's R^2 between the mutual information (MI) of the bivariate marginal distributions in 21 letters and in 4 letters is ≈ 0.995 (Figures S6, S7).

The original MSA was then re-encoded using the reduced per-position alphabet, and the bivariate marginals (Eq. 2) were recalculated using the reduced alphabet. Small pseudocounts are added to the bivariate marginals, as described [32]. Briefly, instead of adding a small flat pseudocount such as $1/N$, we add pseudocounts which correspond to a small per-position chance μ of mutating to a random residue such that the pseudocounted marginals P^{pc} are given by

$$P_{ij}^{pc}(\sigma_i, \sigma_j) = (1-\mu)^2 P_{ij}(\sigma_i, \sigma_j) + \frac{(1-\mu)\mu}{Q}(P_i(\sigma_i) + P_j(\sigma_j)) + \frac{\mu^2}{Q^2} \tag{4}$$

where we take $\mu \approx 1/N$.

Maximum Entropy Model

Following [56], we seek to approximate the unknown empirical probability distribution $P(\vec{\sigma})$ which describes HIV-1 protease sequences $\{\vec{\sigma}\}$ of length L where each residue is encoded in an alphabet of Q states by a model probability distribution $P^m(\vec{\sigma})$. The model distribution we choose is the maximum entropy distribution, e.g. the distribution which maximizes

$$S = -\sum_{k=1}^{Q^L} P^m\left(\vec{\sigma}^k\right) \log P^m\left(\vec{\sigma}^k\right) \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and has been derived by [21, 25, 26, 30, 57] and others satisfying the following constraints:

$$\sum_{k}^{Q^L} P^m\left(\vec{\sigma}^k\right) = 1 \tag{6}$$

$$\sum_{k}^{Q^L} P^m\left(\vec{\sigma}^k\right)\delta\left(\sigma_i^k, \sigma_i\right) = P_i(\sigma_i) \tag{7}$$

$$\sum_{k}^{Q^L} P^m\left(\vec{\sigma}^k\right)\delta\left(\sigma_i^k, \sigma_i\right)\delta\left(\sigma_j^k, \sigma_j\right) = P_{ij}(\sigma_i, \sigma_j) \tag{8}$$

i.e. such that the empirical univariate and bivariate marginal probability distributions are preserved. Through a derivation using Lagrange multipliers not presented here (but can be found in [21, 56]), the maximum entropy model takes the form of a Boltzmann distribution

$$P^m(\vec{\sigma}) = \frac{1}{Z}\exp\left(-\beta E(\vec{\sigma})\right) \tag{9}$$

$$E(\vec{\sigma}) = \sum_{i}^{L} h_i(\sigma_i) + \sum_{i<j}^{L(L-1)/2} J_{ij}(\sigma_i, \sigma_j) \tag{10}$$

where the quantity $E(\vec{\sigma})$ is the Potts Hamiltonian, which determines the statistical energy of a sequence $\vec{\sigma}$, $1/Z$ is a normalization constant, and the inverse temperature $\beta = 1/k_B T$ is such that $k_b T = 1$. This form of the Potts Hamiltonian consists of Lq field parameters h_i and $\binom{L}{2}Q^2$ coupling parameters J_{ij} which describe the system's preference for each amino acid character at site i and each amino acid character pair at sites i, j, respectively. In the way we present the Boltzmann distribution $P^m \propto \exp(-E)$, negative fields and couplings signify favored amino acids preferences.

Not all the model parameters are independent. Due to the relationship between bivariate marginals P_{ij}, P_{ik}, P_{jk} and the fact that the univariate marginals can be derived entirely from the bivariate marginals, only $L(Q-1) + \binom{L}{2}(Q-1)^2$ of these $LQ + \binom{L}{2}Q^2$

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

parameters are independent. Several schemes have been developed and used by others to fully constrain the Hamiltonian (see [25, 26], for example). Further, the fully-constrained Potts Hamiltonian is "gauge invariant" such that the probablity $P^m\left(\vec{\sigma}^k\right)$ is unchanged by (a) a global bias added to the fields, $h_i(\sigma_i) \rightarrow h_i(\sigma_i) + b$, (b) a per-site bias added to the fields $h_i(\sigma_i) \rightarrow h_i(\sigma_i) + b_i$, (c) rearrangement of field and coupling contributions such that $J_{ij}(\sigma_i, \sigma_j) \rightarrow J_{ij}(\sigma_i, \sigma_j) + b_{ij}(\sigma_j)$ and $h_i(\sigma_i) \rightarrow h_i(\sigma_i) - \sum_{j \neq \ i} b_{ij}(\sigma_j)$, or (d) a combination thereof. Due to this gauge invariance, model parameters are over-specified and thus not unique until a fully-constrained gauge is specified, but the properties P^m and ΔE, among others, are gauge invariant and unique among fully-constrained gauges.

Model Inference

Finding a suitable set of Potts parameters $\{h, J\}$ fully determines the total probability distribution $P^m(\vec{\sigma})$ and is achieved by obtaining the set of fields and couplings which yield bivariate marginal estimates $P^m(\sigma_i, \sigma_j)$ that best reproduce the empirical bivariate marginals $P^{obs}(\sigma_i, \sigma_j)$. Previous studies have developed a number of techniques to do this [7, 21, 25, 26, 30, 57–61]. Following [21], we estimate the bivariate marginals given a set of fields and couplings by generating sequences through Markov Chain Monte Carlo (MCMC) where the Metropolis criterion for a generated sequence is proportional to the exponentiated Potts Hamiltonian. The optimal set of parameters $\{h, J\}$ are found through multidimensional Newton search, where bivariate marginal estimates are compared to the empirical distribution to determine descent steps. Unlike several inference methods referenced above, this method avoids making explicit approximations to the model probability distribution, though approximations are made in the computation of the Newton steps, and this method is limited by sampling error of the input emperi-

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cal marginal distributions and by the need for the simulation to equilibrate. Also, the method is computationally intensive. A brief description of the method follows; see the supplemental information of Haldane et al. [32] for a full description of the method.

Determining the schema for choosing the Newton step is crucial. In [21], a quasi-newton parameter update approach was developed, in which updates to J_{ij} and h_i are determined by inverting the system's Jacobian, to minimize the difference between model-estimated and empirical marginals. To simplify and speed up this computation, we take advantage of the gauage invariance of the Potts Hamiltonian to infer a model in which $h_i = 0 \; \forall \; i$, and we compute the expected change in the model marginals $\Delta \; P_{ij}$ (dropping the m superscript) due to a change in J_{ij} to first order by

$$\Delta P_{ij}(\sigma_i, \sigma_j) = \sum_{kl, \sigma_k \sigma_l} \frac{\partial P_{ij}(\sigma_i, \sigma_j)}{\partial J_{kl}(\sigma_k, \sigma_l)} \Delta J_{kl}(\sigma_k, \sigma_l) + \sum_{k, \sigma_k} \frac{\partial P_{ij}(\sigma_i, \sigma_j)}{\partial h_k(\sigma_k)} \Delta h_k(\sigma_k) \tag{11}$$

with a similar relation for $\Delta P_i(\sigma_i)$. The challenge is to compute the Jacobian $\frac{\partial P_{ij}(\sigma_i, \sigma_j)}{\partial J_{kl}(\sigma_k, \sigma_l)}$ and invert the linear system in Equation 11, and solve for the changes ΔJ_{ij} and Δh_i given ΔP_{ij} which we choose as

$$\Delta P_{ij} = \gamma \left(P_{ij}^{emp} - P_{ij} \right) \tag{12}$$

given a damping parameter γ chosen small enough for the linear (and other) approximations to hold.

The computational cost of fitting $\binom{93}{2} \times (4-1)^2 + 93 \times (4-1) = 38,781$ model parameters on 2 NVIDIA K80 or 4 NVIDIA TitanX GPUs is approximately 4 hours. For a more thorough description of the inference methodology, see the supplementary information of Haldane et al. [32].

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Experimental Comparison

Experimentally derived values for either melting temperature (T_m) or viral infectivity via replicative capacity (RC) were mined from the results presented in [9, 40–42]. A csv file of the resulting mined data can be found in Supplementary Data 1.

5 Acknowledgements

This work has been supported in part by grant NIH P50 GM103368 (W.F.F., B.E.T., R.M.L), NIH R01 GM30580 (A.H., R.M.L.), and by computing resource grant NIH S10 OD020095 (W.F.F., A.H., R.M.L.). We thank the supportive collaborative environment provided by the HIV Interaction and Viral Evolution (HIVE) Center at the Scripps Research Institute (`http://hive.scripps.edu`).

References

1. Richman, D. D., Morton, S. C., Wrin, T., Hellmann, N., Berry, S., Shapiro, M. F. & Bozzette, S. A. The prevalence of antiretroviral drug resistance in the United States. *AIDS (London, England)* **18,** 1393–401. ISSN: 0269-9370 (July 2004).

2. Gupta, R., Hill, A., Sawyer, A. W. & Pillay, D. Emergence of drug resistance in HIV type 1-infected patients after receipt of first-line highly active antiretroviral therapy: a systematic review of clinical trials. en. *Clinical infectious diseases* **47,** 712–22. ISSN: 1537-6591. doi:`10.1086/590943` (Sept. 2008).

3. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, N.Y.)* **286,** 295–9. ISSN: 0036-8075. doi:`10.1126/science.286.5438.295` (Oct. 1999).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. en. *Science (New York, N.Y.)* **328,** 1272–5. ISSN: 1095-9203. doi:10.1126/science.1187816 (June 2010).

5. Zeldovich, K. B., Chen, P. & Shakhnovich, E. I. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 16152–7. ISSN: 0027-8424. doi:10.1073/pnas.0705366104 (Oct. 2007).

6. Zeldovich, K. B. & Shakhnovich, E. I. Understanding Protein Evolution: From Protein Physics to Darwinian Selection. *Annual Review of Physical Chemistry* **59,** 105–127. ISSN: 0066-426X. doi:10.1146/annurev.physchem.58.032806.104449 (May 2008).

7. Haq, O., Andrec, M., Morozov, A. V. & Levy, R. M. Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS computational biology* **8,** e1002675. doi:10.1371/journal.pcbi.1002675 (Jan. 2012).

8. Fun, A., Wensing, A. M. J., Verheyen, J. & Nijhuis, M. Human Immunodeficiency Virus Gag and protease: partners in resistance. *Retrovirology* **9,** 63. ISSN: 1742-4690. doi:10.1186/1742-4690-9-63 (Jan. 2012).

9. Chang, M. W. & Torbett, B. E. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *Journal of molecular biology* **410,** 756–60. ISSN: 1089-8638. doi:10.1016/j.jmb.2011.03.038 (July 2011).

10. Flynn, W. F., Chang, M. W., Tan, Z., Oliveira, G., Yuan, J., Okulicz, J. F., Torbett, B. E. & Levy, R. M. Deep Sequencing of Protease Inhibitor Resistant HIV

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Patient Isolates Reveals Patterns of Correlated Mutations in Gag and Protease. *PLoS Comput Biol* **11,** e1004249. doi:10.1371/journal.pcbi.1004249 (2015).

11. Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *The Journal of infectious diseases* **194 Suppl 1,** S51–S58. doi:10.1086/505356 (Sept. 2006).

12. Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J. & Shafer, R. W. Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatments. *Journal of virology* **77,** 4836–4847. doi:10.1128/JVI.77.8.4836-4847.2003 (2003).

13. Shafer, R. W. & Schapiro, J. M. HIV-1 Drug Resistance Mutations: an Updated Framework for the Second Decade of HAART. *AIDS Review* **10,** 67–84 (2008).

14. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of molecular biology* **320,** 85–95. ISSN: 0022-2836. doi:10.1016/S0022-2836(02)00400-X (June 2002).

15. Martinez-Picado, J., Savara, A. V., Sutton, L. & D'Aquila, R. T. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *Journal of virology* **73,** 3744–52. ISSN: 0022-538X (May 1999).

16. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–317. doi:10.1002/prot.340180402 (Apr. 1994).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

17. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. Evolutionary information for specifying a protein fold. *Nature* **437,** 512–8. ISSN: 1476-4687. doi:10.1038/nature03991 (Sept. 2005).

18. Liu, Z., Chen, J. & Thirumalai, D. On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model. *Proteins* **77,** 823–31. ISSN: 1097-0134. doi:10.1002/prot.22498 (Dec. 2009).

19. Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J. & Bonhoeffer, S. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature genetics* **43,** 487–9. ISSN: 1546-1718. doi:10.1038/ng.795 (May 2011).

20. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology* **6,** e1000633. doi:10.1371/journal.pcbi.1000633 (Jan. 2010).

21. Ferguson, A. L., Mann, J. K., Omarjee, S., Ndung'u, T., Walker, B. D. & Chakraborty, A. K. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38,** 606–17. ISSN: 1097-4180. doi:10.1016/j.immuni.2012.11.022 (Mar. 2013).

22. Mann, J. K., Barton, J. P., Ferguson, A. L., Omarjee, S., Walker, B. D., Chakraborty, A. & Ndung'u, T. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. *PLoS Computational Biology* **10** (ed Regoes, R. R.) e1003776. doi:10.1371/journal.pcbi.1003776 (Aug. 2014).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

23. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. en. *Molecular Biology and Evolution* **33,** msv211. ISSN: 0737-4038. doi:10.1093/molbev/msv211 (Oct. 2015).

24. Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Springer, M., Sander, C. & Marks, D. S. Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv: 1510.04612 (Oct. 2015).

25. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 67–72. ISSN: 1091-6490. doi:10.1073/pnas.0805923106 (Jan. 2009).

26. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. & Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108,** E1293–301. ISSN: 1091-6490. doi:10.1073/pnas.1111471108 (Dec. 2011).

27. Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N. & Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 12408–13. ISSN: 1091-6490. doi:10.1073/pnas.1413575111 (Aug. 2014).

28. Sulkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *PNAS* **109,** 10340–10345. doi:10.1073/pnas.1207864109 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

29. Marks, D. S., Hopf, T. a. & Sander, C. Protein structure prediction from sequence variation. *Nature biotechnology* **30,** 1072–80. ISSN: 1546-1696. doi:10.1038/nbt.2419 (Nov. 2012).

30. Barton, J. P., De Leonardis, E., Coucke, A. & Cocco, S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics,* btw328. ISSN: 1367-4803. doi:10.1093/bioinformatics/btw328 (June 2016).

31. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 13567–72. ISSN: 1091-6490. doi:10.1073/pnas.1508584112 (Nov. 2015).

32. Haldane, A., Flynn, W. F., He, P., Vijayan, R. & Levy, R. M. Structural Propensities of Kinase Family Proteins from a Potts Model of Residue Co-Variation. *Protein Science.* ISSN: 09618368. doi:10.1002/pro.2954 (May 2016).

33. Jacquin, H., Gilson, A., Shakhnovich, E., Cocco, S. & Monasson, R. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. *PLOS Computational Biology* **12** (ed Marks, D. S.) e1004889. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1004889 (May 2016).

34. Shekhar, K., Ruberman, C., Ferguson, A., Barton, J., Kardar, M. & Chakraborty, A. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Physical Review E* **88,** 062705. ISSN: 1539-3755. doi:10.1103/PhysRevE.88.062705 (Dec. 2013).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

35. Butler, T. C., Barton, J. P., Kardar, M. & Chakraborty, A. K. Identification of drug resistance mutations in HIV from constraints on natural evolution. *Physical review. E* **93,** 022412. ISSN: 2470-0053. doi:10.1103/PhysRevE.93.022412 (Feb. 2016).

36. Hoffman, N. G., Schiffer, C. A. & Swanstrom, R. Covariation of amino acid positions in HIV-1 protease. *Virology* **314,** 536–548. ISSN: 00426822. doi:10.1016/S0042-6822(03)00484-7 (Sept. 2003).

37. Rhee, S.-Y., Liu, T. F., Holmes, S. P. & Shafer, R. W. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS computational biology* **3,** e87. doi:10.1371/journal.pcbi.0030087 (May 2007).

38. McCandlish, D. M., Otwinowski, J. & Plotkin, J. B. Detecting epistasis from an ensemble of adapting populations. *Evolution; international journal of organic evolution* **69,** 2359–70. ISSN: 1558-5646. doi:10.1111/evo.12735 (Sept. 2015).

39. Gupta, A. & Adami, C. Strong Selection Significantly Increases Epistatic Interactions in the Long-Term Evolution of a Protein. *PLoS genetics* **12,** e1005960. ISSN: 1553-7404. doi:10.1371/journal.pgen.1005960 (Mar. 2016).

40. Muzammil, S., Ross, P. & Freire, E. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. *Biochemistry* **42,** 631–638. ISSN: 00062960. doi:10.1021/bi027019u (Jan. 2003).

41. Louis, J. M., Aniana, A., Weber, I. T. & Sayer, J. M. Inhibition of autoprocessing of natural variants and multidrug resistant mutant precursors of HIV-1 protease by clinical inhibitors. **108,** 9072–7. ISSN: 1091-6490. doi:10.1073/pnas.1102278108 (May 2011).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

42. Henderson, G. J., Lee, S.-K., Irlbeck, D. M., Harris, J., Kline, M., Pollom, E., Parkin, N. & Swanstrom, R. Interplay between single resistance-associated mutations in the HIV-1 protease and viral infectivity, protease activity, and inhibitor sensitivity. *Antimicrobial agents and chemotherapy* **56,** 623–33. ISSN: 1098-6596. doi:10.1128/AAC.05549-11 (Feb. 2012).

43. Van Maarseveen, N. M., de Jong, D., Boucher, C. A. B. & Nijhuis, M. An increase in viral replicative capacity drives the evolution of protease inhibitor-resistant human immunodeficiency virus type 1 in the absence of drugs. *Journal of acquired immune deficiency syndromes* **42,** 162–8. ISSN: 1525-4135. doi:10.1097/01.qai.0000219787.65915.56 (June 2006).

44. Ode, H., Neya, S., Hata, M., Sugiura, W. & Hoshino, T. Computational Simulations of HIV-1 ProteasesMulti-drug Resistance Due to Nonactive Site Mutation L90M. *Journal of the American Chemical Society* **128,** 7887–7895. doi:10.1021/ja060682b (2006).

45. Mahalingam, B., Wang, Y.-F., Boross, P. I., Tozser, J., Louis, J. M., Harrison, R. W. & Weber, I. T. Crystal structures of HIV protease V82A and L90M mutants reveal changes in the indinavir-binding site. *European journal of biochemistry / FEBS* **271,** 1516–24. ISSN: 0014-2956. doi:10.1111/j.1432-1033.2004.04060.x (Apr. 2004).

46. Kovalevsky, A. Y., Tie, Y., Liu, F., Boross, P. I., Wang, Y. F., Leshchenko, S., Ghosh, A. K., Harrison, R. W. & Weber, I. T. Effectiveness of nonpeptide clinical inhibitor TMC-114 on HIV-1 protease with highly drug resistant mutations D30N, I50V, and L90M. *Journal of Medicinal Chemistry* **49,** 1379–1387. ISSN: 00222623. doi:10.1021/jm050943c (Feb. 2006).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

47. Kouyos, R. D., Leventhal, G. E., Hinkley, T., Haddad, M., Whitcomb, J. M., Petropoulos, C. J. & Bonhoeffer, S. Exploring the complexity of the HIV-1 fitness landscape. *PLoS genetics* **8,** e1002551. ISSN: 1553-7404. doi:10.1371/journal.pgen.1002551 (Jan. 2012).

48. Poelwijk, F. J., Tănase-Nicola, S., Kiviet, D. J. & Tans, S. J. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology* **272,** 141–4. ISSN: 1095-8541. doi:10.1016/j.jtbi.2010.12.015 (Mar. 2011).

49. Kolli, M., Stawiski, E., Chappey, C. & Schiffer, C. A. Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance. *Journal of virology* **83,** 11027–42. ISSN: 1098-5514. doi:10.1128/JVI.00628-09 (Nov. 2009).

50. Parry, C. M., Kolli, M., Myers, R. E., Cane, P. A., Schiffer, C. A. & Pillay, D. Three residues in HIV-1 matrix contribute to protease inhibitor susceptibility and replication capacity. *Antimicrobial agents and chemotherapy* **55,** 1106–13. ISSN: 1098-6596. doi:10.1128/AAC.01228-10 (Mar. 2011).

51. Prabu-Jeyabalan, M., Nalivaika, E. & Schiffer, C. a. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **10,** 369–81. ISSN: 0969-2126. doi:10.1016/S0969-2126(02)00720-7 (Mar. 2002).

52. Breuer, S., Sepulveda, H., Chen, Y., Trotter, J. & Torbett, B. E. A cleavage enzyme-cytometric bead array provides biochemical profiling of resistance mutations in HIV-1 Gag and protease. *Biochemistry* **50,** 4371–4381. ISSN: 00062960. doi:10.1021/bi200031m (2011).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

53. Johnson, V. a., Calvez, V., Gunthard, H. F., Paredes, R., Pillay, D., Shafer, R. W., Wensing, A. M. & Richman, D. D. Update of the drug resistance mutations in HIV-1: March 2013. *Topics in antiviral medicine* **21,** 6–14. ISSN: 2161-5853 (2013).

54. Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., Wei, X., Jiang, C., Kirchherr, J. L., Gao, F., Anderson, J. A., Ping, L.-H., Swanstrom, R., Tomaras, G. D., Blattner, W. A., Goepfert, P. A., *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 7552–7. ISSN: 1091-6490. doi:10.1073/pnas.0802203105 (May 2008).

55. Murphy, L. R., Wallqvist, A. & Levy, R. M. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering* **13,** 149–152. doi:10.1093/protein/13.3.149 (2000).

56. Mora, T. & Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **144,** 268–302. ISSN: 0022-4715. doi:10.1007/s10955-011-0229-4 (June 2011).

57. Mézard, M. & Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of physiology, Paris* **103,** 107–13. ISSN: 1769-7115. doi:10.1016/j.jphysparis.2009.05.013 (2009).

58. Cocco, S. & Monasson, R. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. *Physical Review Letters* **106,** 090601. ISSN: 0031-9007. doi:10.1103/PhysRevLett.106.090601 (Mar. 2011).

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

59. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)* **28,** 184–90. ISSN: 1367-4811. doi:10.1093/bioinformatics/btr638 (Jan. 2012).

60. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* **87,** 012707. ISSN: 1539-3755. doi:10.1103/PhysRevE.87.012707 (Jan. 2013).

61. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79,** 1061–78. ISSN: 1097-0134. doi:10.1002/prot.22934.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures and Figure Legends

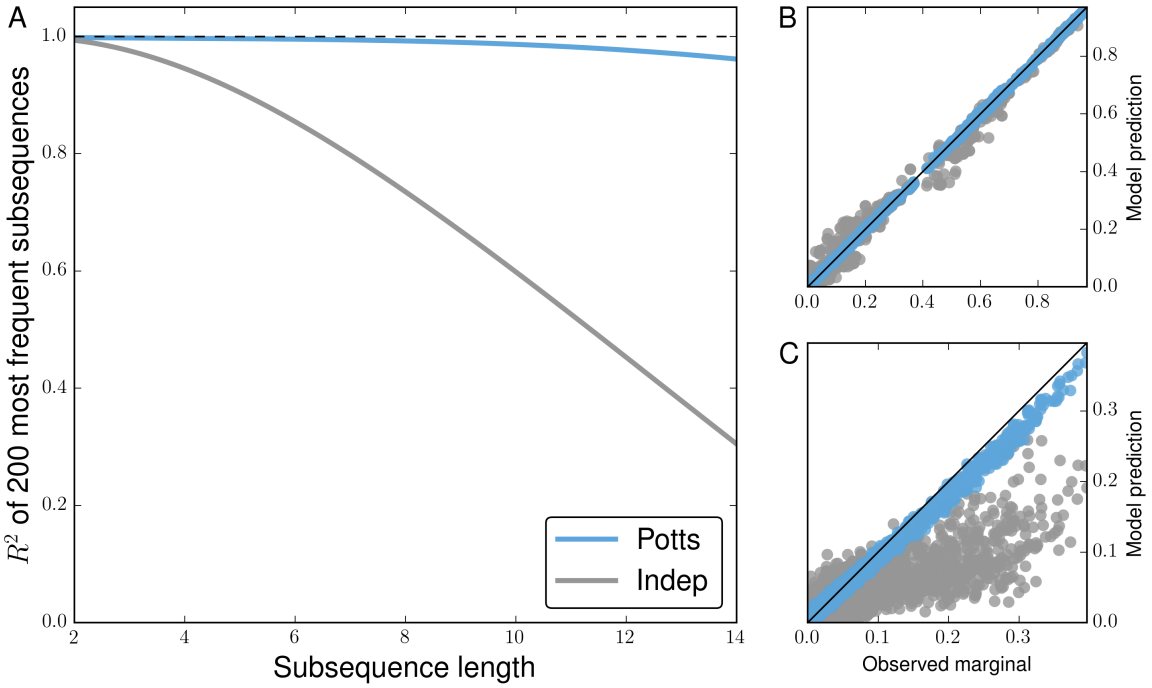


Figure 1: **Potts model is predictive of higher order sequence statistics.** For each subsequence length varying from 2 to 14, subsequence frequencies are computed for all observed subsequences at 500 randomly chosen combinations among 36 PI-associated positions. (A) Pearson R^2 of the 200 most probable observed subsequence frequencies (marginals) with corresponding predictions by Potts (blue) and independent (gray) models for varying subsequence lengths. (B) 2^{nd} and (C) 14^{th} order observed marginals predicted by both models. Shown in (B,C) are observed frequencies at the 500 randomly chosen combinations of 2 and 14 positions among 36 PI-associated sites, with approximately 2500 and 5600 subsequence frequencies greater than 0.01 visible, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

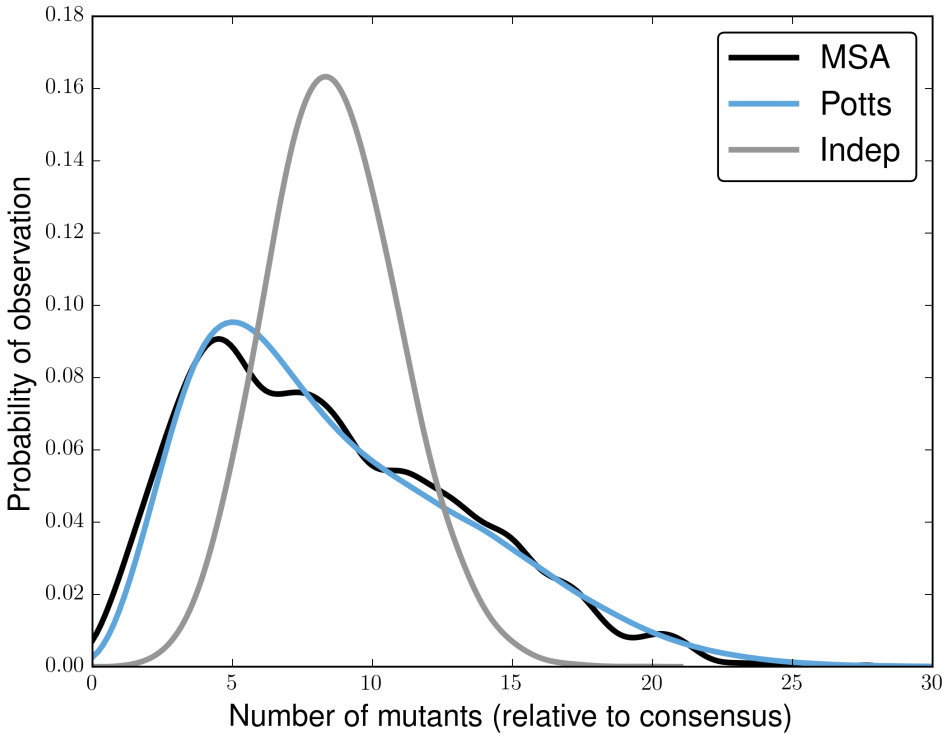


Figure 2: **Potts model captures properties of full length sequence ensemble.** Probabilities of observing sequences with any k mutations relative to the consensus sequence as observed in original MSA (black) and predicted by the Potts (blue) and independent (gray) models.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

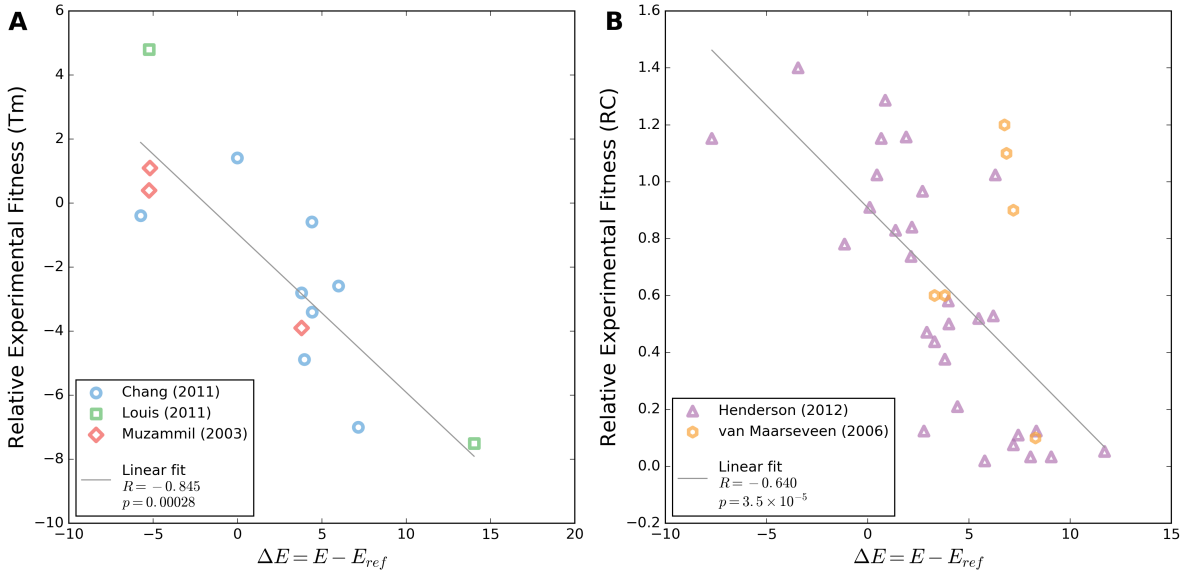


Figure 3: **Change in Potts energy correlates with change in experimental fitness.** (A) Changes in melting temperature (T_m) for individual sequences relative to a reference sequence extracted from literature [9, 40, 41]. These sequences differ from the wildtype by 1–2 mutations [9] up to 10–14 mutations [40, 41]. (B) Change in relative infectivty as measured by replicative capacity assay for individual sequences containing only single point mutations [42] and 1–5 mutations [43]. In both panels a linear regression fit with Pearson's R and associated two-tailed p-value are provided in the legend.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

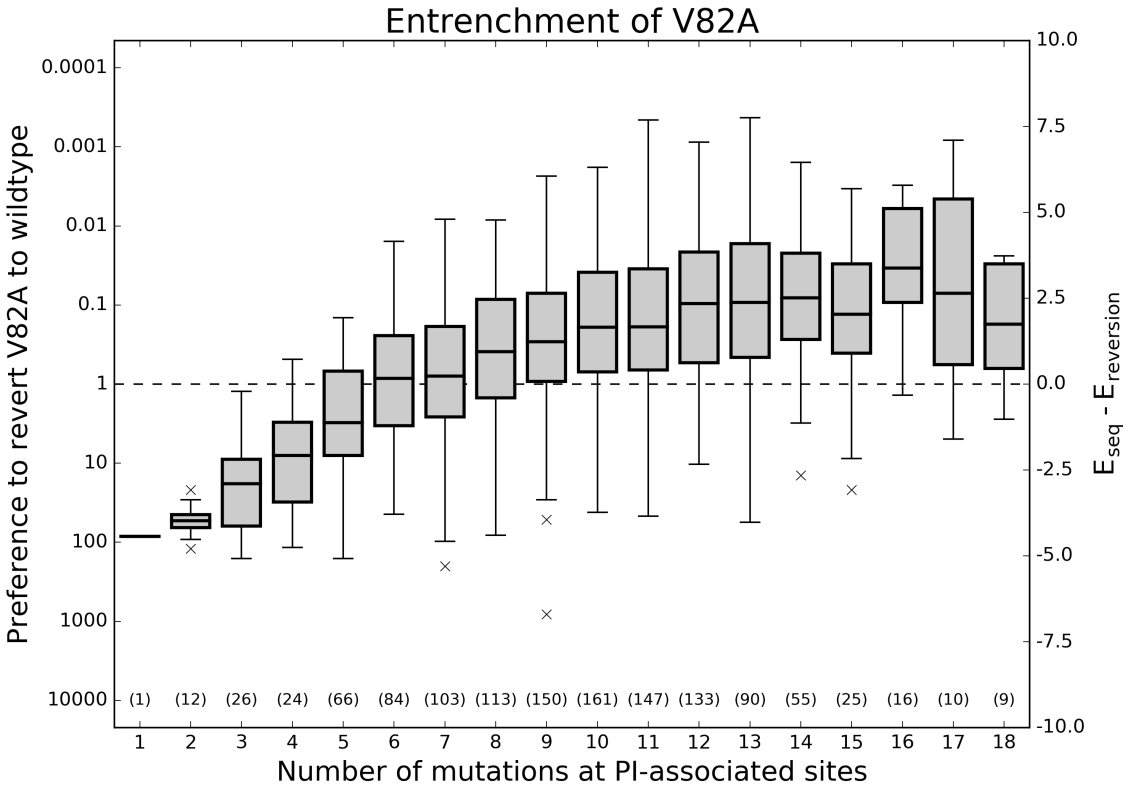
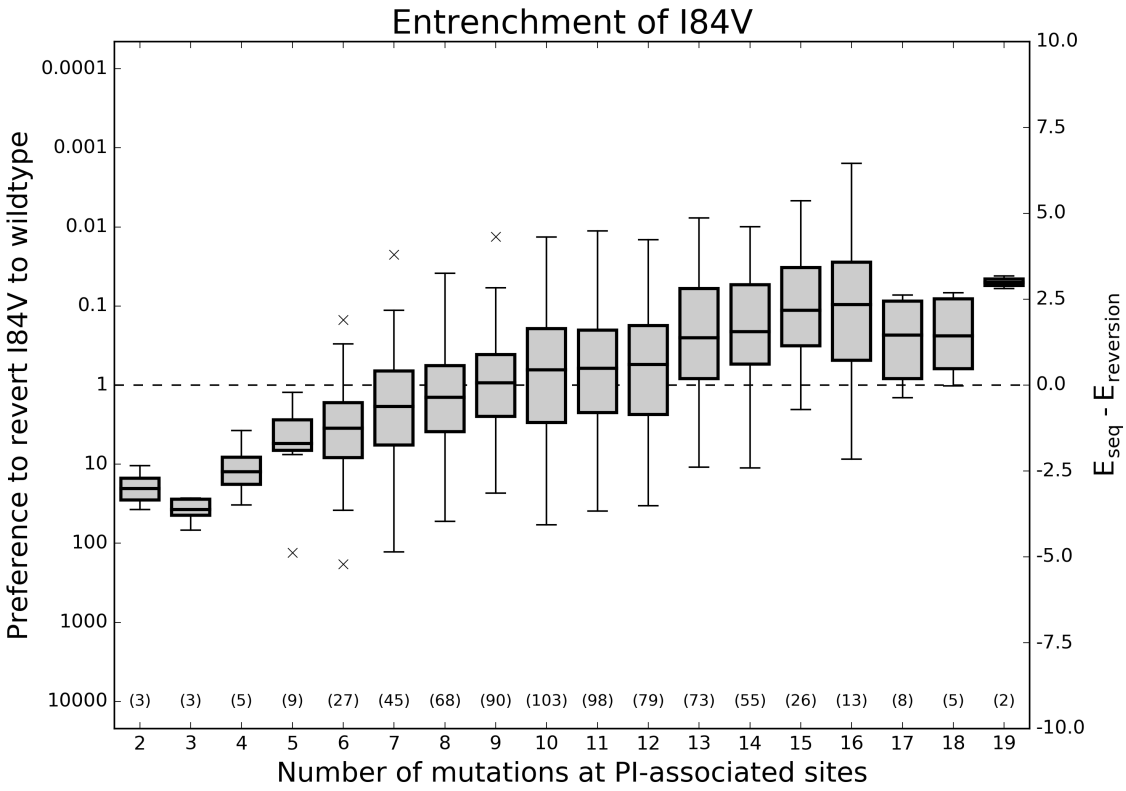
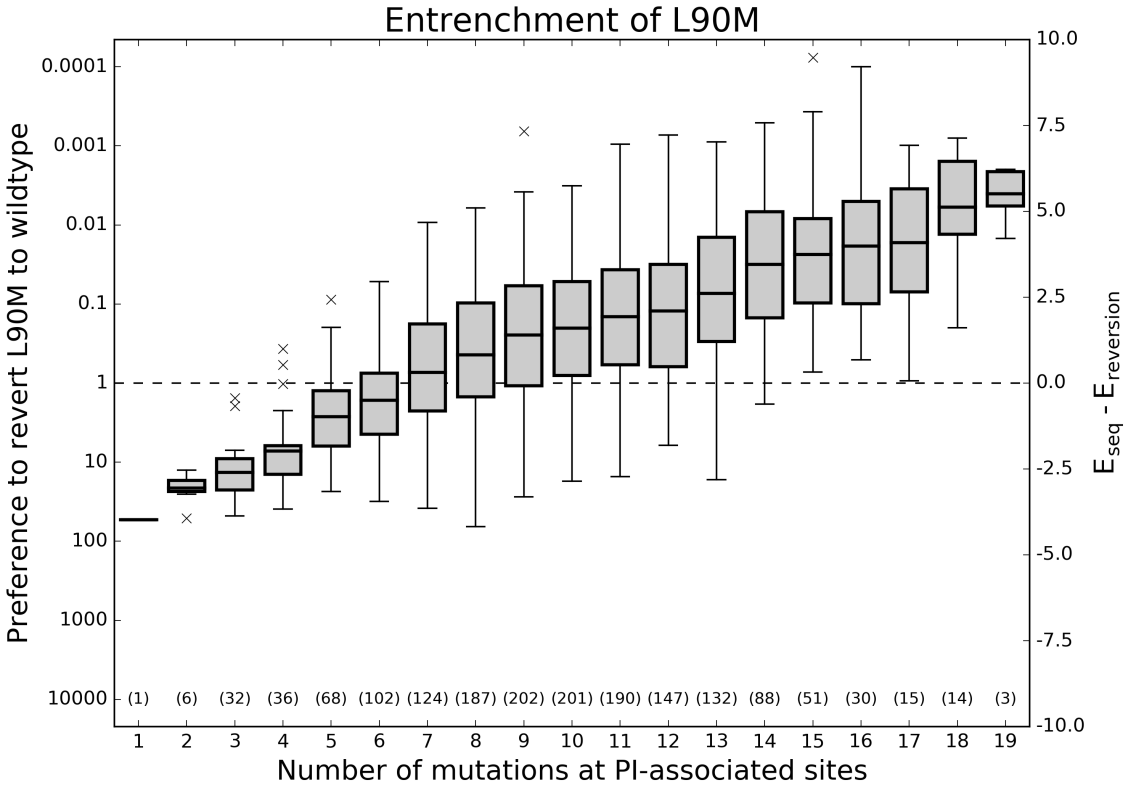


Figure 4: **Effect of epistasis on the fitness penalty incurred by primary resistance mutations.** For each of the 3 primary HIV protease mutations described in [9], two Potts statistical energies are computed for all observed sequences containing that mutation: E_{seq}, the energy of the sequence with that mutation and $E_{reversion}$, the energy with that primary mutation reverted to wildtype. This Potts energy difference, $\Delta E = E_{seq} - E_{reversion}$ is shown versus hamming distance from the wildtype including only PI-assocated positions. Ordinate scales are given in both relative probability of reversion $\exp(-\Delta E)$ (left) and ΔE (right). Values below (above) the dashed line on the ordinate correspond to fitness gain (penalty) upon reversion to wildtype. Although primary resistance mutations initially destabilize the protease, as mutations accumulate, the primary resistance mutations become entrenched, meaning their reversion becomes destabilizing to the protein.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.





bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figures

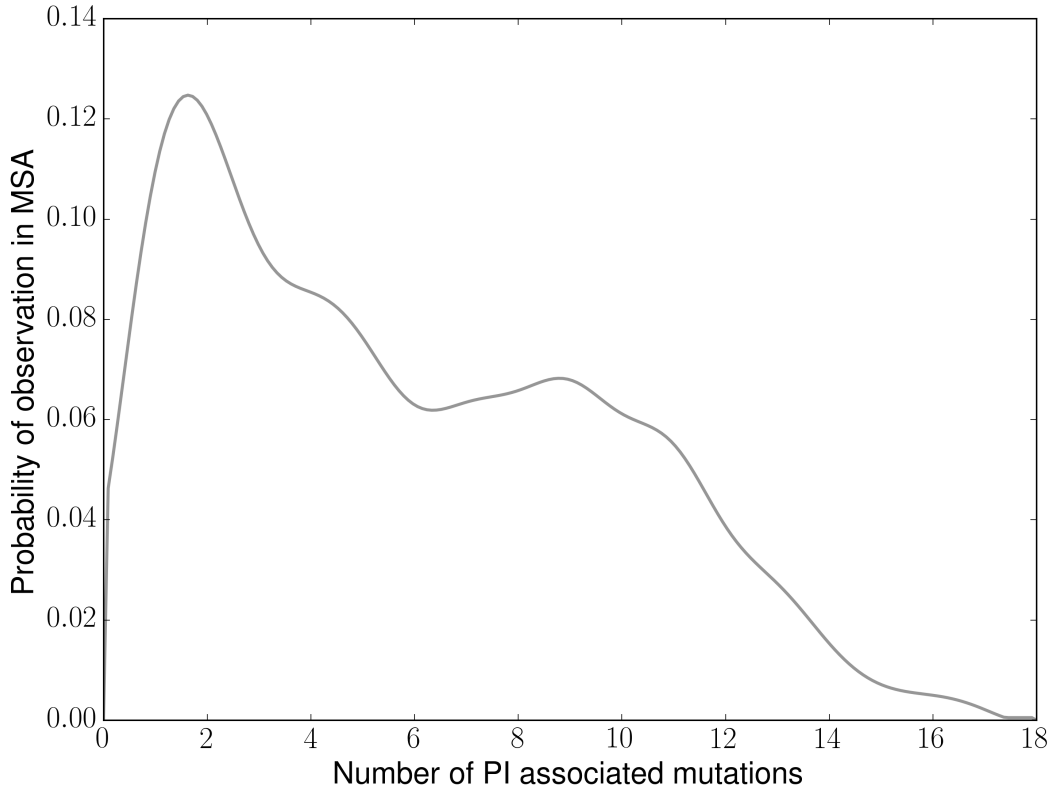


Figure S1: **Probabilities of observing sequences with k PI-associated mutations in our PI-experienced dataset from the Stanford HIVDB.** Mutations considered PI-associated were extracted from [53] and include: L10I/F/V/C/R, V11I, G16E, K20R/M/I/T/V, L24I, D30N, V32I, L33I/F/V, E34Q, M36I/L/V, K43T, M46I/L, I47V/A, G48V, I50L/V, F53L/Y, I54V/L/A/M/T/S, Q58E, D60E, I62V, L63P, I64L/M/V, H69K/R, A71V/I/T/L, G73S/A/C/T, T74P, L76V, V77I, V82A/F/T/S/L/I, N83D, I84V, I85V, N88D/S, L89I/M/V, L90M, I93L/M.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

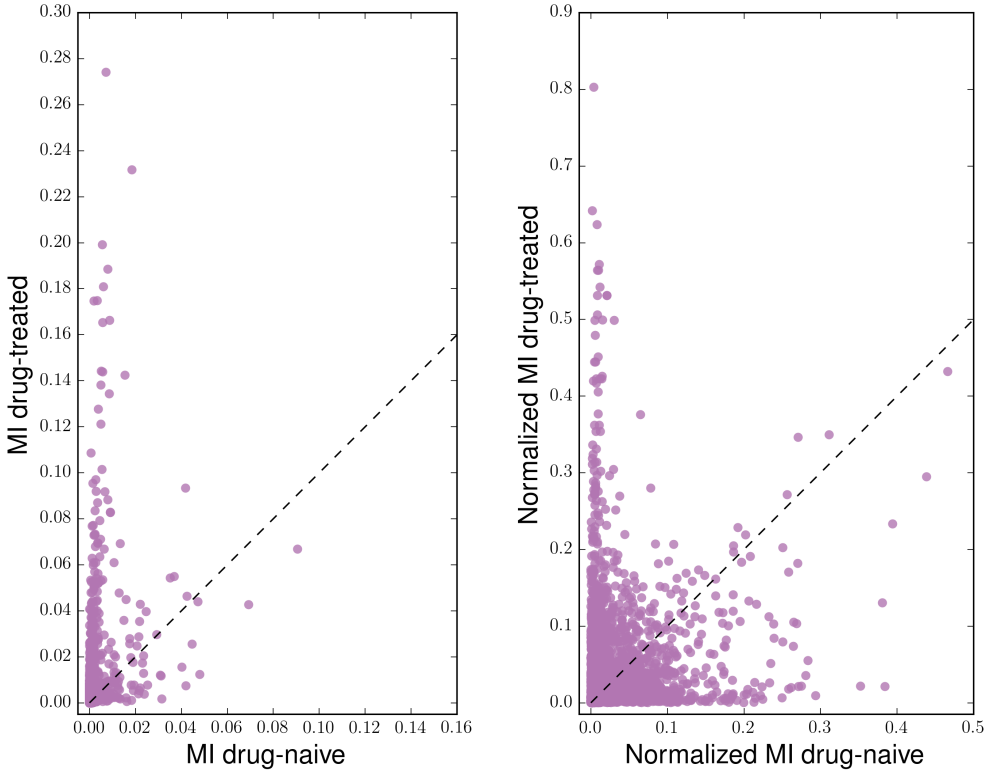


Figure S2: **Mutual information and normalized mutual information for drug-experienced and drug-naive sequences.** Correlated information for each pair of positions in drug-experienced and drug-naive HIV protease sequences determined using mutual information (MI) and a normalized variant of MI assuming mutual information a special case of the total correlation (TC). TC is a multivariate generalization of mutual information, and for the relevant case of pair marginals its maximum takes the form $TC_{ij}^{\max} = \min(H(P_i), H(P_j))$, where $H(P) = -\sum_k P(k) \log P(k)$ is the Shannon entropy. (left) MI_{ij} and (right) $MI_{ij}/TC_{ij}^{\max}$ measured in bits for all observed pair marginals in drug-experienced and drug-naive sequences. Drug-experienced sequences exhibit correlations several times larger in magnitude than those in drug-naive sequences, even when normalized by the information content constrained on the univariate marginals.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

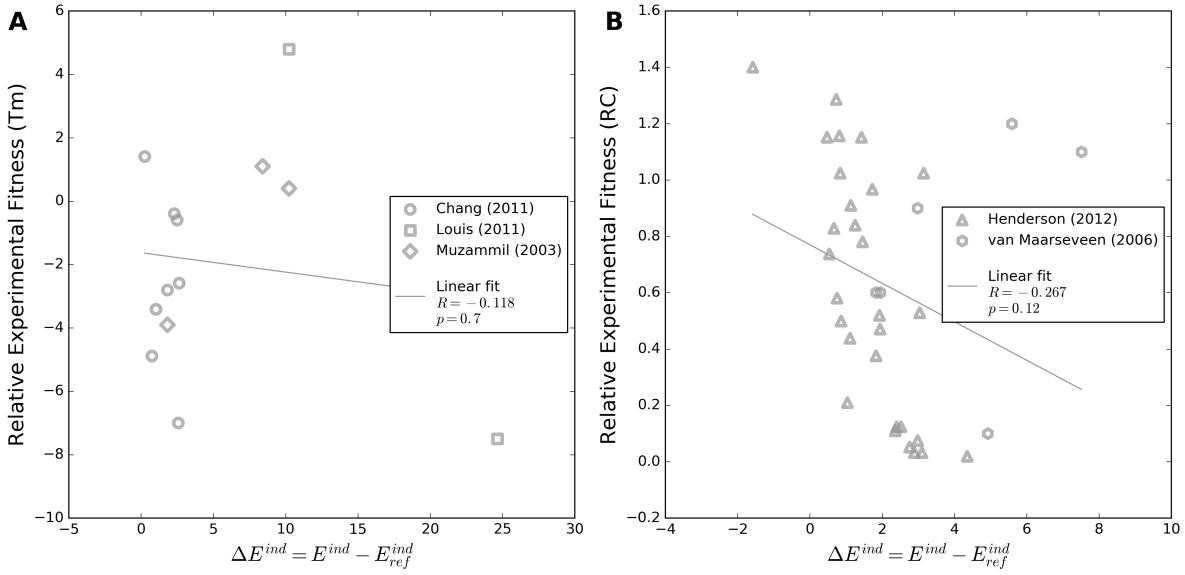


Figure S3: **Change in independent model energy does not correlates with change in experimental fitness.** (A) Changes in melting temperature (T_m) and (B) relative infectivity by replicative capacity assay for individual sequences relative to a reference sequence extracted from literature as shown in Figure 3[9, 40–43]. In both panels a linear regression fit with Pearson's R and associated two-tailed p-value are provided in the legend.

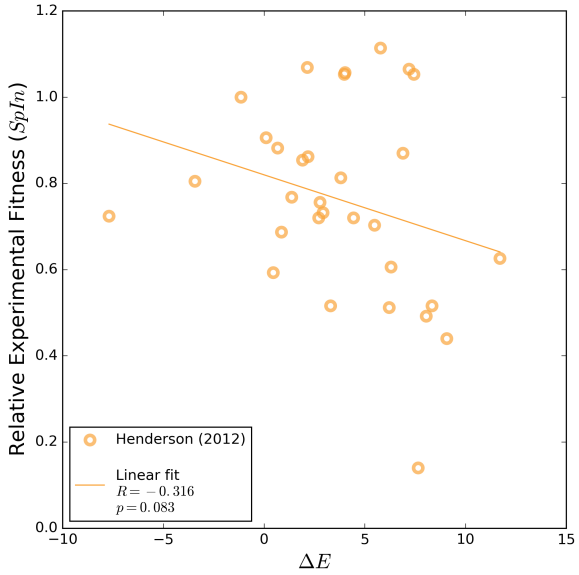
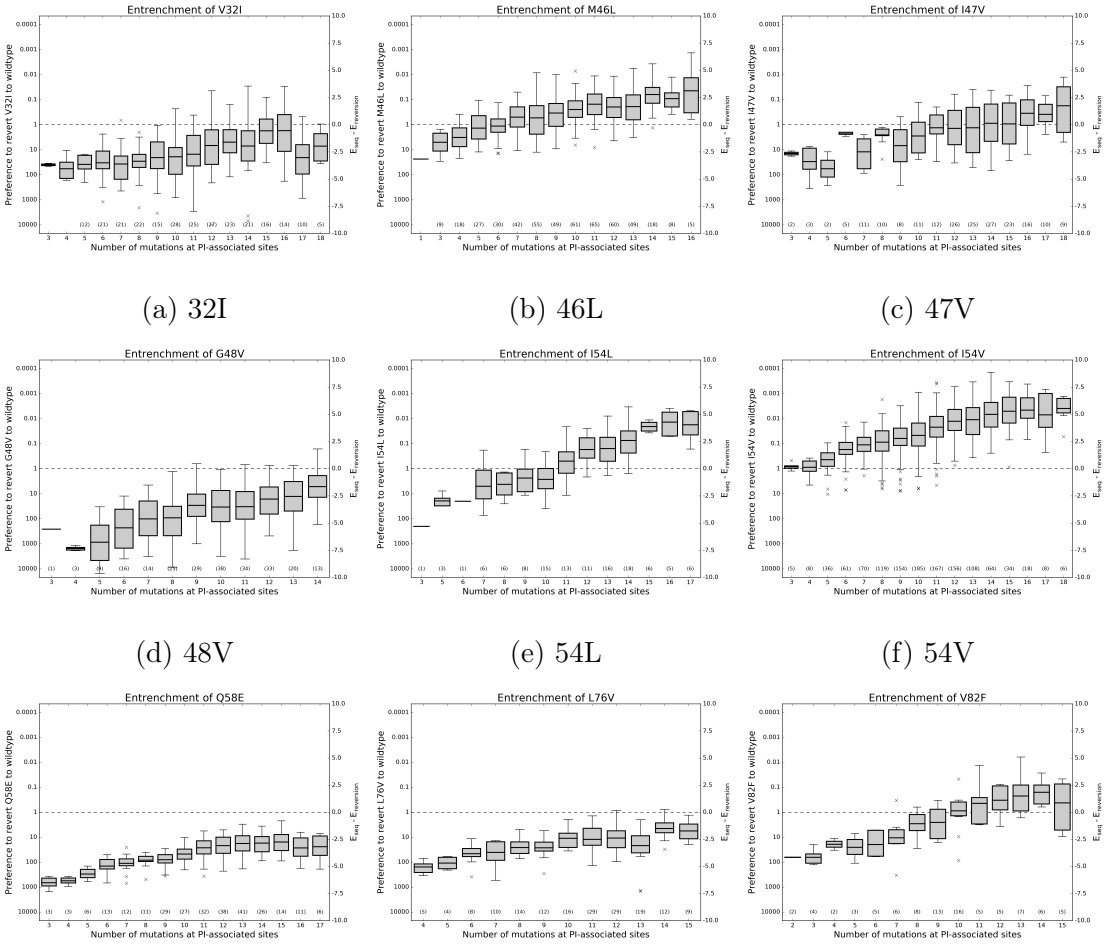


Figure S4: **Additional experimental comparision of Potts model statistical energies.** Relative infectivity by SpIn assay for individual single mutant sequences relative to a reference sequence extracted from [42]. Linear regression fits with Pearson's R and associated two-tailed p-value are provided.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) 32I (b) 46L (c) 47V

(d) 48V (e) 54L (f) 54V

(g) 58E (h) 76V (i) 82F

Figure S5: Entrenchment for a selection of primary and primary/accessory resistance mutations. Each shows a similar trend of increasing mean $\Delta E_{reversion}$ shown in Figure 4 for primary mutations V82A, I84V, and L90M, meaning the mutations become less destabilizing on average as background mutations accumulate, although not all mutations shown here cross from destabilizing to stabilizing. Note that for some mutations the number of observed sequences with that mutation may be small (≤ 10) for some values of hamming distance from wildtype.

bioRxiv preprint doi: https://doi.org/10.1101/063750; this version posted July 13, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

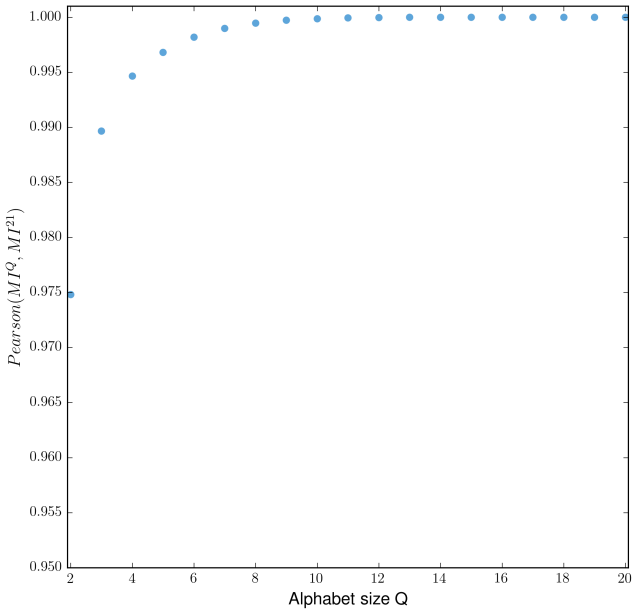


Figure S6: Pearson R^2 of the mutual information (MI) of bivariate marginals of each position pair in the 21 letter alphabet and Q letter alphabet as Q is varied.

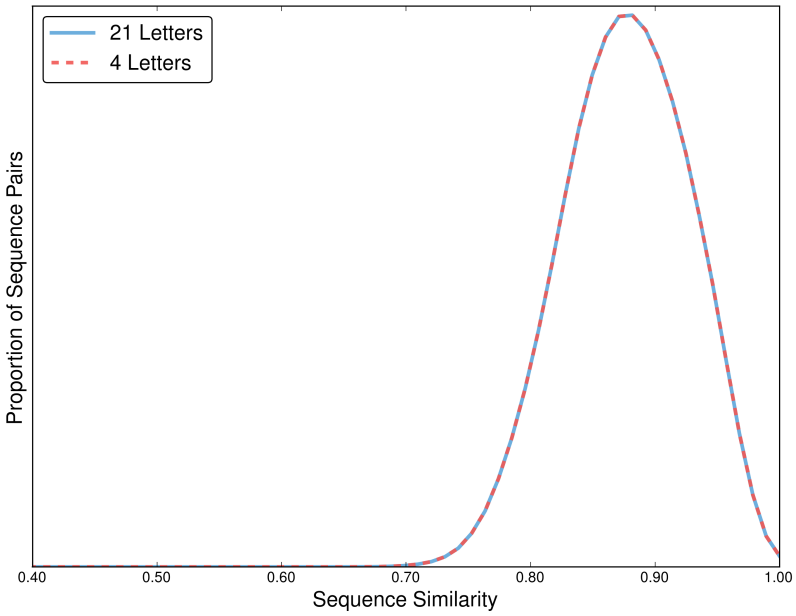


Figure S7: The distribution of sequence similarities in the 21 letter alphabet (blue) and 4 letter alphabet (dashed red).

```
RefID,PMID,Author,Year,Nmut,Mut,Base,Method,NiceName,Res,Err
1,21762813,Chang,2011,1,I84V,Wild,DSC,Tm,−2.803,
1,21762813,Chang,2011,1,V82A,Wild,DSC,Tm,−3.41,
1,21762813,Chang,2011,1,L90M,Wild,DSC,Tm,−4.885,
1,21762813,Chang,2011,2,I84V/L90M,Wild,DSC,Tm,−7,
1,21762813,Chang,2011,2,L10I/I84V,Wild,DSC,Tm,−0.393,
1,21762813,Chang,2011,2,L63P/I84V,Wild,DSC,Tm,1.41,
1,21762813,Chang,2011,2,A71V/I84V,Wild,DSC,Tm,−0.59,
1,21762813,Chang,2011,2,V77I/I84V,Wild,DSC,Tm,−2.59,
2,12534275,Muzammil,2003,1,I84V,Wild,DSC,Tm,−3.9,
2,12534275,Muzammil,2003,11,L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/I84V/L90M/I93L,Wil
d,DSC,Tm,0.4,
2,12534275,Muzammil,2003,10,L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/L90M/I93L,Wild,DSC
,Tm,1.1,
3,10452615,Xie,1999,3,Q7K/L33I/L63I,D25N,Kd_dimer,K_d,5.8,2.1
3,10452615,Xie,1999,1,V82F,D25N,Kd_dimer,K_d,134,12
3,10452615,Xie,1999,2,V82F/I84V,D25N,Kd_dimer,K_d,35.9,7.4
3,10452615,Xie,1999,2,V82T/I84V,D25N,Kd_dimer,K_d,82,11
3,10452615,Xie,1999,1,L90M,D25N,Kd_dimer,K_d,29.2,4
4,8621402,Szeltner,1996,3,Q7K/L33I/L63I,Wild,dG,,1.38,0.71
6,22083488,Henderson,2012,1,I15V,Wild,ELISA,SpIn,1,
6,22083488,Henderson,2012,1,E35D,Wild,ELISA,SpIn,1.057,
6,22083488,Henderson,2012,1,N37D,Wild,ELISA,SpIn,0.862,
6,22083488,Henderson,2012,1,I64V,Wild,ELISA,SpIn,0.882,
6,22083488,Henderson,2012,1,L10I,Wild,ELISA,SpIn,0.724,
6,22083488,Henderson,2012,1,M36I,Wild,ELISA,SpIn,0.593,
6,22083488,Henderson,2012,1,I62V,Wild,ELISA,SpIn,0.687,
6,22083488,Henderson,2012,1,L63P,Wild,ELISA,SpIn,0.805,
6,22083488,Henderson,2012,1,A71V,Wild,ELISA,SpIn,0.768,
6,22083488,Henderson,2012,1,A71T,Wild,ELISA,SpIn,0.72,
6,22083488,Henderson,2012,1,V77I,Wild,ELISA,SpIn,0.854,
6,22083488,Henderson,2012,1,I93L,Wild,ELISA,SpIn,1.069,
6,22083488,Henderson,2012,1,K20R,Wild,ELISA,SpIn,0.703,
6,22083488,Henderson,2012,1,K20I,Wild,ELISA,SpIn,0.756,
6,22083488,Henderson,2012,1,L24I,Wild,ELISA,SpIn,0.626,
6,22083488,Henderson,2012,1,D30N,Wild,ELISA,SpIn,0.516,
6,22083488,Henderson,2012,1,V32I,Wild,ELISA,SpIn,0.492,
6,22083488,Henderson,2012,1,M46I,Wild,ELISA,SpIn,0.906,
6,22083488,Henderson,2012,1,M46L,Wild,ELISA,SpIn,0.732,
6,22083488,Henderson,2012,1,I47V,Wild,ELISA,SpIn,0.606,
6,22083488,Henderson,2012,1,G48V,Wild,ELISA,SpIn,0.44,
6,22083488,Henderson,2012,1,I50V,Wild,ELISA,SpIn,0.14,
6,22083488,Henderson,2012,1,F53L,Wild,ELISA,SpIn,0.512,
6,22083488,Henderson,2012,1,I54V,Wild,ELISA,SpIn,0.516,
6,22083488,Henderson,2012,1,G73S,Wild,ELISA,SpIn,1.053,
6,22083488,Henderson,2012,1,V82A,Wild,ELISA,SpIn,0.72,
6,22083488,Henderson,2012,1,V82T,Wild,ELISA,SpIn,1.065,
6,22083488,Henderson,2012,1,I84V,Wild,ELISA,SpIn,0.813,
6,22083488,Henderson,2012,1,N88D,Wild,ELISA,SpIn,0.87,
6,22083488,Henderson,2012,1,N88S,Wild,ELISA,SpIn,1.114,
6,22083488,Henderson,2012,1,L90M,Wild,ELISA,SpIn,1.053,
6,22083488,Henderson,2012,1,I15V,Wild,RTPCR,SpIn,0.776,
6,22083488,Henderson,2012,1,E35D,Wild,RTPCR,SpIn,0.744,
6,22083488,Henderson,2012,1,N37D,Wild,RTPCR,SpIn,0.545,
6,22083488,Henderson,2012,1,I64V,Wild,RTPCR,SpIn,0.89,
6,22083488,Henderson,2012,1,L10I,Wild,RTPCR,SpIn,0.821,
6,22083488,Henderson,2012,1,M36I,Wild,RTPCR,SpIn,0.467,
6,22083488,Henderson,2012,1,I62V,Wild,RTPCR,SpIn,0.7,
6,22083488,Henderson,2012,1,L63P,Wild,RTPCR,SpIn,0.87,
6,22083488,Henderson,2012,1,A71V,Wild,RTPCR,SpIn,0.654,
6,22083488,Henderson,2012,1,A71T,Wild,RTPCR,SpIn,0.972,
6,22083488,Henderson,2012,1,V77I,Wild,RTPCR,SpIn,0.732,
6,22083488,Henderson,2012,1,I93L,Wild,RTPCR,SpIn,0.927,
6,22083488,Henderson,2012,1,K20R,Wild,RTPCR,SpIn,0.728,
6,22083488,Henderson,2012,1,K20I,Wild,RTPCR,SpIn,0.72,
6,22083488,Henderson,2012,1,L24I,Wild,RTPCR,SpIn,0.83,
6,22083488,Henderson,2012,1,D30N,Wild,RTPCR,SpIn,0.411,
6,22083488,Henderson,2012,1,V32I,Wild,RTPCR,SpIn,0.821,
6,22083488,Henderson,2012,1,M46I,Wild,RTPCR,SpIn,0.679,
```

```
6,22083488,Henderson,2012,1,M46L,Wild,RTPCR,SpIn,0.573,
6,22083488,Henderson,2012,1,I47V,Wild,RTPCR,SpIn,0.663,
6,22083488,Henderson,2012,1,G48V,Wild,RTPCR,SpIn,0.42,
6,22083488,Henderson,2012,1,I50V,Wild,RTPCR,SpIn,0.118,
6,22083488,Henderson,2012,1,F53L,Wild,RTPCR,SpIn,0.374,
6,22083488,Henderson,2012,1,I54V,Wild,RTPCR,SpIn,0.463,
6,22083488,Henderson,2012,1,G73S,Wild,RTPCR,SpIn,0.768,
6,22083488,Henderson,2012,1,V82A,Wild,RTPCR,SpIn,0.626,
6,22083488,Henderson,2012,1,V82T,Wild,RTPCR,SpIn,0.862,
6,22083488,Henderson,2012,1,I84V,Wild,RTPCR,SpIn,0.683,
6,22083488,Henderson,2012,1,N88D,Wild,RTPCR,SpIn,1.028,
6,22083488,Henderson,2012,1,N88S,Wild,RTPCR,SpIn,1,
6,22083488,Henderson,2012,1,L90M,Wild,RTPCR,SpIn,1.053,
6,22083488,Henderson,2012,1,I15V,Wild,RC,RC,0.781,
6,22083488,Henderson,2012,1,E35D,Wild,RC,RC,0.5,
6,22083488,Henderson,2012,1,N37D,Wild,RC,RC,0.84,
6,22083488,Henderson,2012,1,I64V,Wild,RC,RC,1.152,
6,22083488,Henderson,2012,1,L10I,Wild,RC,RC,1.152,
6,22083488,Henderson,2012,1,M36I,Wild,RC,RC,1.024,
6,22083488,Henderson,2012,1,I62V,Wild,RC,RC,1.286,
6,22083488,Henderson,2012,1,L63P,Wild,RC,RC,1.4,
6,22083488,Henderson,2012,1,A71V,Wild,RC,RC,0.829,
6,22083488,Henderson,2012,1,A71T,Wild,RC,RC,0.967,
6,22083488,Henderson,2012,1,V77I,Wild,RC,RC,1.157,
6,22083488,Henderson,2012,1,I93L,Wild,RC,RC,0.738,
6,22083488,Henderson,2012,1,K20R,Wild,RC,RC,0.52,
6,22083488,Henderson,2012,1,K20I,Wild,RC,RC,0.124,
6,22083488,Henderson,2012,1,L24I,Wild,RC,RC,0.052,
6,22083488,Henderson,2012,1,D30N,Wild,RC,RC,0.124,
6,22083488,Henderson,2012,1,V32I,Wild,RC,RC,0.033,
6,22083488,Henderson,2012,1,M46I,Wild,RC,RC,0.91,
6,22083488,Henderson,2012,1,M46L,Wild,RC,RC,0.471,
6,22083488,Henderson,2012,1,I47V,Wild,RC,RC,1.024,
6,22083488,Henderson,2012,1,G48V,Wild,RC,RC,0.033,
6,22083488,Henderson,2012,1,I50V,Wild,RC,RC,,
6,22083488,Henderson,2012,1,F53L,Wild,RC,RC,0.529,
6,22083488,Henderson,2012,1,I54V,Wild,RC,RC,0.438,
6,22083488,Henderson,2012,1,G73S,Wild,RC,RC,0.11,
6,22083488,Henderson,2012,1,V82A,Wild,RC,RC,0.21,
6,22083488,Henderson,2012,1,V82T,Wild,RC,RC,0.076,
6,22083488,Henderson,2012,1,I84V,Wild,RC,RC,0.376,
6,22083488,Henderson,2012,1,N88D,Wild,RC,RC,,
6,22083488,Henderson,2012,1,N88S,Wild,RC,RC,0.019,
6,22083488,Henderson,2012,1,L90M,Wild,RC,RC,0.581,
6,22083488,Henderson,2012,1,I15V,Wild,EC50,EC50,1.16,0.15
6,22083488,Henderson,2012,1,E35D,Wild,EC50,EC50,0.59,0.08
6,22083488,Henderson,2012,1,N37D,Wild,EC50,EC50,0.71,0.1
6,22083488,Henderson,2012,1,I64V,Wild,EC50,EC50,1.04,0.13
6,22083488,Henderson,2012,1,L10I,Wild,EC50,EC50,1.16,0.13
6,22083488,Henderson,2012,1,M36I,Wild,EC50,EC50,0.84,0.12
6,22083488,Henderson,2012,1,I62V,Wild,EC50,EC50,0.84,0.15
6,22083488,Henderson,2012,1,L63P,Wild,EC50,EC50,1.26,0.16
6,22083488,Henderson,2012,1,A71V,Wild,EC50,EC50,1.83,0.39
6,22083488,Henderson,2012,1,A71T,Wild,EC50,EC50,1.16,0.23
6,22083488,Henderson,2012,1,V77I,Wild,EC50,EC50,0.92,0.12
6,22083488,Henderson,2012,1,I93L,Wild,EC50,EC50,0.62,0.16
6,22083488,Henderson,2012,1,K20R,Wild,EC50,EC50,0.62,0.07
6,22083488,Henderson,2012,1,K20I,Wild,EC50,EC50,0.47,0.09
6,22083488,Henderson,2012,1,L24I,Wild,EC50,EC50,0.64,0.15
6,22083488,Henderson,2012,1,D30N,Wild,EC50,EC50,0.44,0.15
6,22083488,Henderson,2012,1,V32I,Wild,EC50,EC50,1.04,0.6
6,22083488,Henderson,2012,1,M46I,Wild,EC50,EC50,1.17,0.17
6,22083488,Henderson,2012,1,M46L,Wild,EC50,EC50,0.75,0.11
6,22083488,Henderson,2012,1,I47V,Wild,EC50,EC50,1.43,0.54
6,22083488,Henderson,2012,1,G48V,Wild,EC50,EC50,2.51,0.6
6,22083488,Henderson,2012,1,I50V,Wild,EC50,EC50,1.08,1.04
6,22083488,Henderson,2012,1,F53L,Wild,EC50,EC50,0.68,0.23
6,22083488,Henderson,2012,1,I54V,Wild,EC50,EC50,0.51,0.14
6,22083488,Henderson,2012,1,G73S,Wild,EC50,EC50,0.52,0.12
```

```
6,22083488,Henderson,2012,1,V82A,Wild,EC50,EC50,0.46,0.12
6,22083488,Henderson,2012,1,V82T,Wild,EC50,EC50,0.63,0.27
6,22083488,Henderson,2012,1,I84V,Wild,EC50,EC50,0.8,0.26
6,22083488,Henderson,2012,1,N88D,Wild,EC50,EC50,0.54,0.26
6,22083488,Henderson,2012,1,N88S,Wild,EC50,EC50,0.78,0.5
6,22083488,Henderson,2012,1,L90M,Wild,EC50,EC50,0.72,0.14
7,14622012,Ohtaka,2003,6,L10I/M46I/I54V/V82A/I84V/L90M,Wild,Kcat,,0.081,
7,14622012,Ohtaka,2003,4,M46I/I54V/V82A/I84V,Wild,Kcat,,0.0844,
7,14622012,Ohtaka,2003,2,V82A/I84V,Wild,Kcat,,0.3966,
7,14622012,Ohtaka,2003,2,M46I/I54V,Wild,Kcat,,0.345,
7,14622012,Ohtaka,2003,2,L10I/L90M,Wild,Kcat,,0.147,
8,10196268,Martinez-Picado,1999,1,D30N,Wild,TCID/p24,TCID/p24,0.29,
8,10196268,Martinez-Picado,1999,2,D30N/L63P,Wild,TCID/p24,TCID/p24,1.15,
8,10196268,Martinez-Picado,1999,1,L90M,Wild,TCID/p24,TCID/p24,0.72,
8,10196268,Martinez-Picado,1999,2,L63P/L90M,Wild,TCID/p24,TCID/p24,0.79,
8,10196268,Martinez-Picado,1999,5,L10R/M46I/L63P/V82T/I84V,Wild,TCID/p24,TCID/p24,1.3,
10,21576495,Louis,2011,11,L10I/M36I/S37D/M46I/R57K/L63P/A71V/G73S/I84V/L90M/I93L,Wild,
DSC,Tm,4.8,
10,21576495,Louis,2011,14,L10I/I15V/K20R/L24I/V32I/L33F/M36I/M46L/I54M/L63P/K70Q/V82I/
I84V/L89M,Wild,DSC,Tm,-7.5,
11,16645546,van Maarseveen,2006,1,I84V,Wild,RC,RC,0.6,
11,16645546,van Maarseveen,2006,2,M36I/I54V,Wild,RC,RC,0.6,
11,16645546,van Maarseveen,2006,1,V82T,Wild,RC,RC,0.9,
11,16645546,van Maarseveen,2006,3,M36I/I54V/V82T,Wild,RC,RC,0.1,
11,16645546,van Maarseveen,2006,4,M36I/I54V/A71V/V82T,Wild,RC,RC,1.2,
11,16645546,van Maarseveen,2006,5,K20R/M36I/I54V/A71V/V82T,Wild,RC,RC,1.1,
```