Supporting Information

I. INFERRING THE ENERGY FUNCTION

We downloaded a multiple sequence alignment (MSA) for the HIV-1 clade B Protease protein from the Los Alamos National Laboratory HIV database (http://www.hiv.lanl.gov). Sequences labeled by the database as "problematic" were excluded. To minimize evolved drug resistance we only selected sequences obtained in the year 1996 or earlier, and we removed sequences from trial studies of protease inhibitors as described in the main text, yielding a total of 6701 sequences from 757 unique patients. After downloading, the MSA data was processed to remove insertions relative to the HXB2 reference sequence [1]. Ambiguous amino acids were then imputed with simple mean imputation.

We determined the most common amino acid at each position in the protein, which we refer to as the "wild-type" amino acid. We then translated each sequence in the MSA into a binary form by assigning a 0 to each position where the amino acid matched the wild-type, and a 1 to each position where there was a mismatch, as described in the main text. Protease is highly conserved: the consensus amino acid was observed in a super-majority ($\geq$80%) of the sequence data at 94% of sites. We thus expect that a binary representation of the data will be sufficient to capture useful information about correlated mutations in these proteins.

The binarized MSA data consists of sequences from B patients, which we label $k = 1, \ldots, B$. Let us call the number of sequences from the kth patient as B_k, and let us write the ath sequence from patient k as $s^{(k,a)} = \{s_1^{(k,a)}, \ldots, s_{99}^{(k,a)}\}$, with the single site variables $s_i \in \{0, 1\}$. To obtain a represenative sample of the population we averaged over multiple sequences from the same patient, so the one- and two-point correlations we obtain from the data are then

$$p_i = \frac{1}{B}\sum_{k=1}^{B}\left[\frac{1}{B_k}\sum_{a=1}^{B_k} s_i^{(k,a)}\right], \qquad p_{ij} = \frac{1}{B}\sum_{k=1}^{B}\left[\frac{1}{B_k}\sum_{a=1}^{B_k} s_i^{(k,a)} s_j^{(k,a)}\right]. \tag{1}$$

The one-point correlations p_i measure the frequency of mutations at each position i, and the two-point correlations p_{ij} measure the frequency of pairs of mutations occuring simultaneously at two positions i, j.

Our goal is to infer a probability distribution which reproduces the empirical correlations. It can be shown that the least constrained (maximum entropy) model capable of reproducing the correlations is an Ising model with pairwise interactions, wherein the probability of observing a sequence is

$$P(s) = \frac{e^{-E(s)}}{Z}, \quad E(s) = \sum_{i=1}^{L} h_i s_i + \sum_{i<j} J_{ij} s_i s_j, \tag{2}$$

with the partition function $Z = \sum_s \exp(-E(s))$. The parameters $\{h_i\}, \{J_{ij}\}$ must then be chosen so that the correlations obtained from the Ising model match the empirical correlations,

$$\langle s_i \rangle = \frac{1}{Z}\sum_s s_i\, e^{-E(s)} = p_i, \quad \langle s_i s_j \rangle = \frac{1}{Z}\sum_s s_i s_j\, e^{-E(s)} = p_{ij}. \tag{3}$$

Note that the sums in Eq. 3 are over all 2^L binary sequences of length L.

The difficult computational problem of solving Eq. 3 is referred to as the inverse Ising problem. It is possible to show [2] that the $\{h_i\}, \{J_{ij}\}$ satisfying Eq. 3 are those which minimize

$$\log Z(\{h_i\}, \{J_{ij}\}) + \sum_{i=1}^{L} h_i p_i + \sum_{i<j} J_{ij} p_{ij}. \tag{4}$$

However, no analytical solution exists for the parameters minimizing Eq. 4, and numerical approaches are precluded for systems with $L \gtrsim 20$ because the number of operations necessary to compute the partition function Z is exponential in L. To solve this problem we employ the selective cluster expansion (SCE) method [2–4] introduced by Cocco and Monasson, which constructs an estimate for the proper $\{h_i\}$, $\{J_{ij}\}$ by directly solving Eq. 4 for small subsets of the full system and combining the results. For thorough reviews of this method and example applications, see [2, 4].

II. PREDICTING HIGHER ORDER MOMENTS OF THE PROBABILITY DISTRIBUTION

Unlike the $\{p_i\}$ and $\{p_{ij}\}$, higher order statistics, such as three-point correlations or the probability $P(n)$ of observing sequences with n mutations with respect to the wild-type sequence, are not constrained in the inference

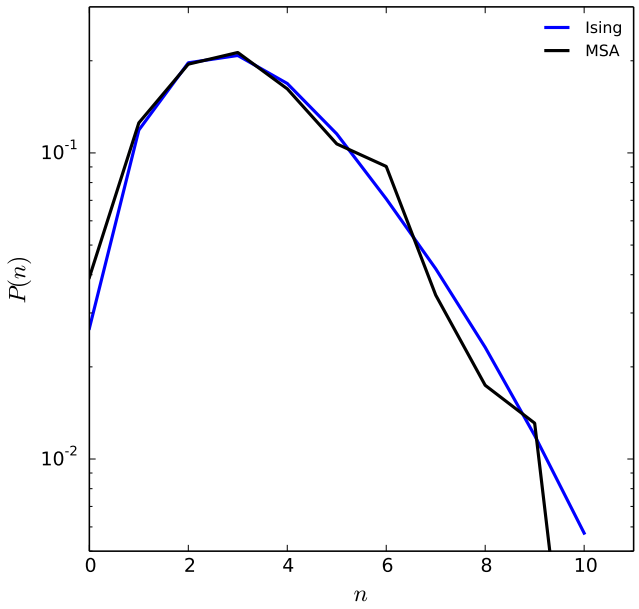

FIG. S1: Comparison of the probability of observing a sequence with n mutations in the MSA and in the inferred Ising model.

problem. A good fit to these higher order statistics is thus a measure of the *predictive power* of the inferred Ising model. Because mutations in the protease protein are rare at most sites, typical three-point correlations are small, and thus more sensitive to noise due to finite sampling. To evaluate the model's prediction of higher order moments of the probability distribution, we thus compare the model $P(n)$ curve versus that obtained from the MSA. As shown in Fig. S1, Ising model predictions for the $P(n)$ curve match very well with MSA data.

III. SINGLE SITE APPROXIMATION

To motivate the use of the two site approximation in the main text, we show here that the values of the one-point correlations $\{s_i\}$ are largely determined within a single site approximation (i.e. an approximation that neglects all couplings between sites). To establish this, we show that solving for the correlations within the single site approximation captures 90 percent of the variance (R^2) of the correlations when solved for without a single site approximation.

Within the single site approximation, the inference problem Eq. 3 reduces to

$$\langle s_i \rangle = \frac{\exp(-h_i)}{1+\exp(-h_i)} \tag{5}$$

for each site. A comparison of the results for the one-point correlations using Eq. 5 to the frequency of single-site mutations in the MSA is shown in Fig. S2.

IV. PREDICTING THE VALUES OF THE PARAMETERS $\{h_i^f\}$

Here we demonstrate that the claim in the main text that the fitness values of the fields $\{h_i^f\}$ are difficult to infer from the values of $\{h_i\}$ inferred from the observed correlations. With the Eigen model in Ising form [5] in a single

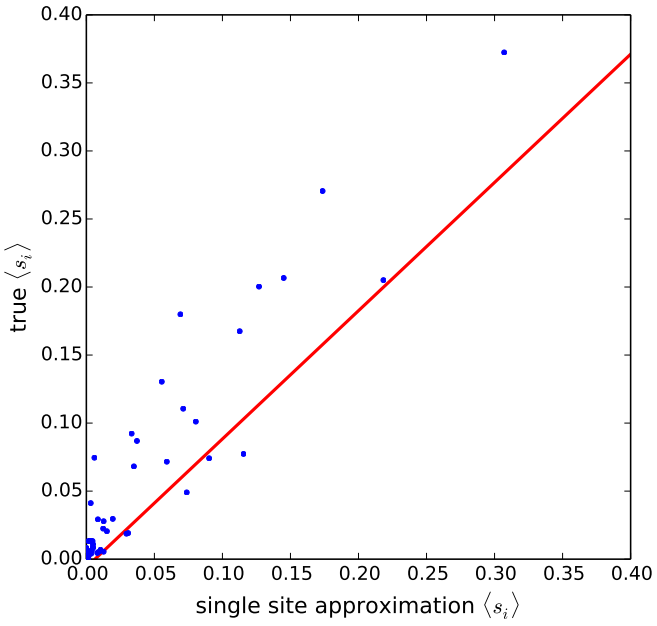

FIG. S2: Linear regression of the single site approximation solution for the one-point correlations $\langle s_i \rangle$ against the frequency of single-site mutations in the MSA. Here $R^2 \approx 0.89$.

site approximation,

$$\exp(-E(s^T)) \propto \sum_{\{s^t\}_{t=0}^{T-1}} \exp\left[\sum_{t=0}^{T-1} K(2s_t - 1)(2s_{t+1} - 1) - F(s_t))\right]$$

$$F(s) = \sum_{i=0}^{L} h_i^f s_i, \tag{6}$$

we can solve for each site by decomposing the sum in Eq. 6 into a product of transfer matrices

$$M = \begin{pmatrix} \exp(K-h) & \exp(-K) \\ \exp(-K-h) & \exp(K) \end{pmatrix}. \tag{7}$$

In the limit of many generations, we can rewrite Eq. 6 as

$$\exp(-E(s^T)) \propto \lim_{T\to\infty} M^T v^0, \tag{8}$$

where v^0 is a vector with the proportion of the population initially in the wild type and mutant states. This implies that we can obtain all of the information about the asymptotic state by looking at the eigenvector associated with the largest eigenvalue of M. Solving for h in the very small h^f limit yields

$$h = \exp(2K)\, h^f + O((h^f)^3). \tag{9}$$

The exponential dependence on K, and the value of $K \simeq 4$ for amino acid mutations in HIV leads to extreme sensitivity in h to small changes in h^f for small h^f, and the slow change in h for larger changes in h^f observed for larger h^f make inferring h^f from h a very difficult problem in practice. However, it is likely that these issues are moderated at population sizes that are finite.

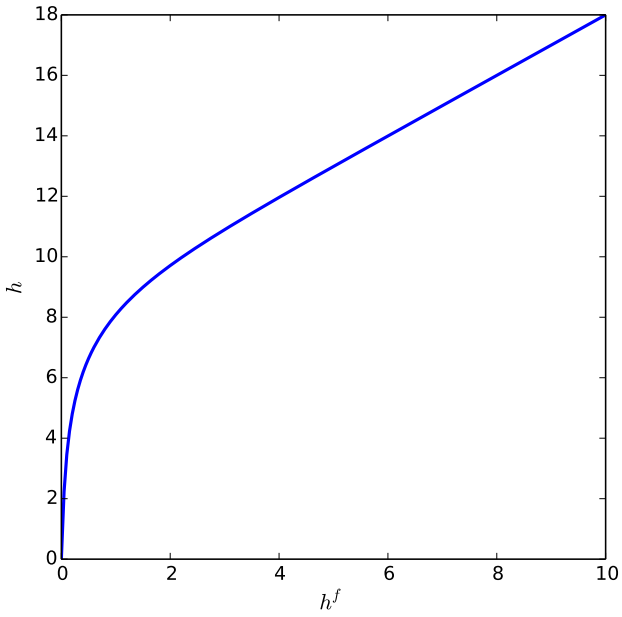

FIG. S3: Plot of h versus h^f showing the sensitivity of h. The inferred field h approaches h^f as $h^f \to \infty$.

V. TRANSFER MATRIX FOR THE TWO SITE MODEL

To compute the solution for the Eigen model in the two site approximation, the following transfer matrix was used:

$$M = \begin{pmatrix} \exp(2K - h_1 - h_2 - J) & \exp(-h_1) & \exp(-h_2) & \exp(-2K) \\ \exp(-h_1 - h_2 - J) & \exp(2K - h_1) & \exp(-2K - h_2) & 1 \\ \exp(-h_1 - h_2 - J) & \exp(-2K - h_1) & \exp(2K - h_2) & 1 \\ \exp(-2K - h_1 - h_2 - J) & \exp(-h_1) & \exp(-h_2) & \exp(2K) \end{pmatrix}. \tag{10}$$

The normalized elements of the eigenvector associated with the largest eigenvalue give the fraction of the population in each state, and are trivially algebraically related to the parameters of the prevalence landscape.

VI. STATISTICS OF COUPLINGS FOR STABILIZING ACCESSORY MUTATIONS

We checked if the couplings between the major resistance site 84, and the associated accessory mutation sites 10, 63, and 71 [6], are larger than would be expected randomly. To check this, we computed the probability that the average coupling of the three sites with site 84 can be generated by choosing three random pairs of sites. The resulting p-value is $\simeq 0.0518$. The individual coupling values are $J_{10,84} = 1.16$, $J_{63,84} = 1.04$, and $J_{71,84} = 0.27$ (average value 0.82), which lie in the top 6th, 7th, and 13th percentile of all couplings, respectively.

VII. STATISTICAL SIGNIFICANCE OF RESISTANCE MUTATION DETECTION

As a further check of the significance of the results, we computed p-values for the null hypothesis that predicted resistance sites were drawn randomly. This results in a p-value that is a function of number of predicted resistance sites. If there are r sites randomly drawn out of a total of $N = 99$ sites, and m of the sites drawn are resistance sites (out of $M = 12$), the p-value is given by

$$p = \sum_{k=m}^{M} \frac{\binom{M}{k}\binom{N-M}{r-k}}{\binom{N}{r}} \tag{11}$$

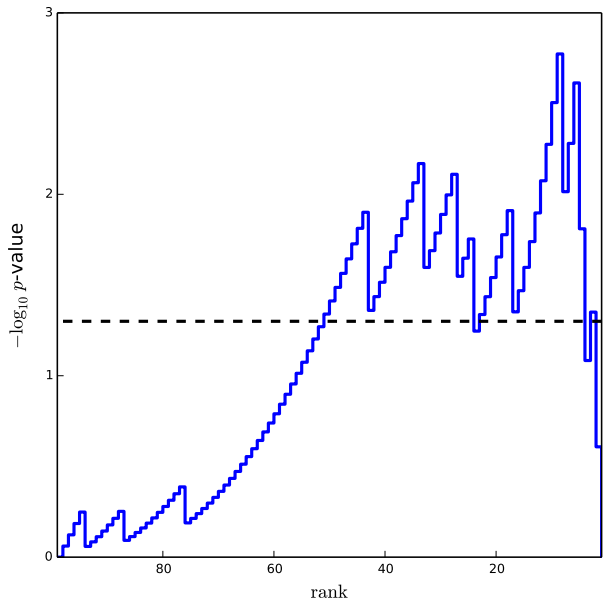

FIG. S4: Minus log p-values (base 10) as a function of rank. The dashed black line indicates the standard significance threshold 0.05.

The p-values are plotted in Fig. S4 as a function of rank r. As noted in the main text, $p < 0.05$ for almost all ranks between 3 and 50, supporting the significance of the results, as the classification rule is not expected to perform well for weakly coupled sites (low ranks).

VIII. RESULTS FOR ALTERNATIVE CLASSIFICATION PROCEDURES AND DRUG NAÏVE DATA

Here we show predictions of resistance sites using alternative classification rules and data. We first examine the predictions made with the same model, but including all sequences from drug-naïve patients up until the present. The results are shown in Fig. S5, along with the calculation from the main text for comparison, and are not significantly different. This is probably because transmitted protease inhibitor resistance is relatively rare [7, 8]. However, the performance is slightly better at the extremely high threshold limit for the drug-naïve sequence case, a possible signature of transmitted drug resistance.

Another very simple way to make predictions is to simply threshold the observed correlation matrix, defined by

$$C_{ij} = \frac{\langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle}{\sqrt{\langle s_i \rangle (1 - \langle s_i \rangle) \langle s_j \rangle (1 - \langle s_j \rangle)}} \,. \tag{12}$$

In principle, all of the arguments developed in the main text apply to correlations as well. However, the presumed advantage of the direct interactions approach is that it disentangles indirect from direct interactions, which the correlation matrix does not. Predictions using the correlation matrix compared with the direct interactions approach (with all sequences from drug-naïve patients, as well as the restricted sequence set used in the main text) are in Fig. S5. The direct interaction approach clearly performs better for the high ranked sites.

In protein contact prediction, a common measure of interactions is the direct information. Direct information is defined with respect to a two site model

$$P(s_i, s_j) = Z^{-1} \exp\left(J_{ij} s_i s_j + \tilde{h}_i s_i + \tilde{h}_j s_i\right). \tag{13}$$

The coupling J_{ij} is taken from the full solution of the inverse Ising problem with all sites, and the fields $\tilde{h}_i$ and $\tilde{h}_j$ are chosen to match the single site probabilities $P(s_i)$ and $P(s_j)$. The direct information between sites i and j is then

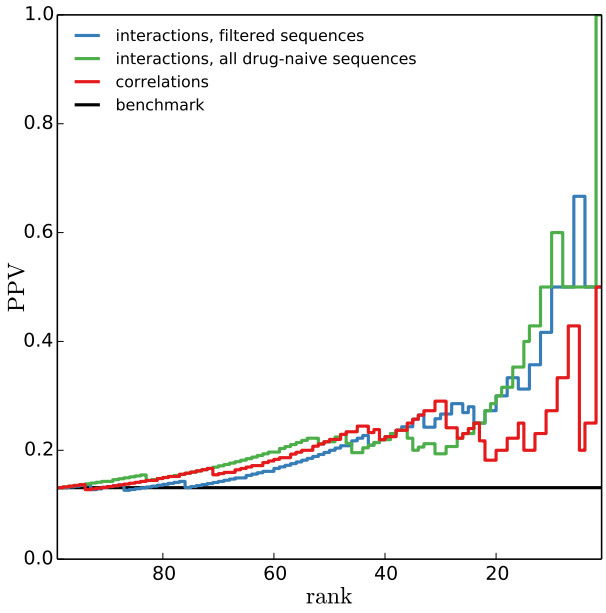

FIG. S5: Comparison of classification results for positive predictive value (PPV).

constructed as

$$DI_{ij} = \sum_{s_i,s_j} P(s_i,s_j) \log\left(\frac{P(s_i,s_j)}{P(s_i)P(s_j)}\right). \tag{14}$$

It is important to note that the direct information only contains contributions from the inferred direct interactions between sites, and not any information about the network. In this sense, it is distinct from mutual information, which contains network effects. Thresholding the direct information matrix, and following the usual procedure for predicting resistance mutations results in predictions of resistance sites. The results are shown in Fig. S6.

We note also that many of the largest couplings link sites where just one site is classified as a major site of drug resistance. Based on the methods presented here, we have no way to distinguish which site or sites in a strongly linked pair should be associated with drug resistance. One alternate approach, then, would be to rank the couplings in order of their strength and attempt to predict how often either one or both coupled sites are sites of major drug resistance. Performance on this classification problem is also substantially better than random for the largest couplings, as shown in Fig. S7.

IX. THE RELATIONSHIP BETWEEN PROTEIN CONTACTS AND PREDICTED RESISTANCE SITES

The methods used in this paper are closely related to methods used for predicting protein contacts [9, 10]. In the context of the protein contact prediction problem, large $-J_{ij}$ values would be associated with pairs of contact residues, not resistance sites. However, the large couplings here are only weakly correlated with contact sites. This can be shown clearly in a protein contact map with predicted resistance mutations shown on the matrix with cross coordinates connected to their closest neighbour resistance mutations. To create this map (Fig. S8), we considered residues with alpha carbons within 8 Å to be in contact [9], and considered the entire homodimer of protease, rather than a single monomer in order to include contacts due to the dimeric structure. The structure used was the structure 1A30 from the protein database [11], which has been used for previous studies of protease inhibitors and evolution [12].

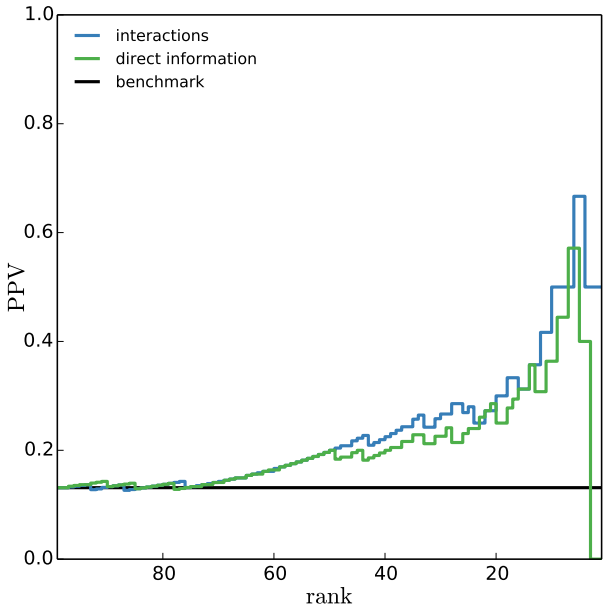

FIG. S6: Comparison of direct information approach and the direct interaction approach from the paper to classifying drug resistance mutations using positive predictive value (PPV).

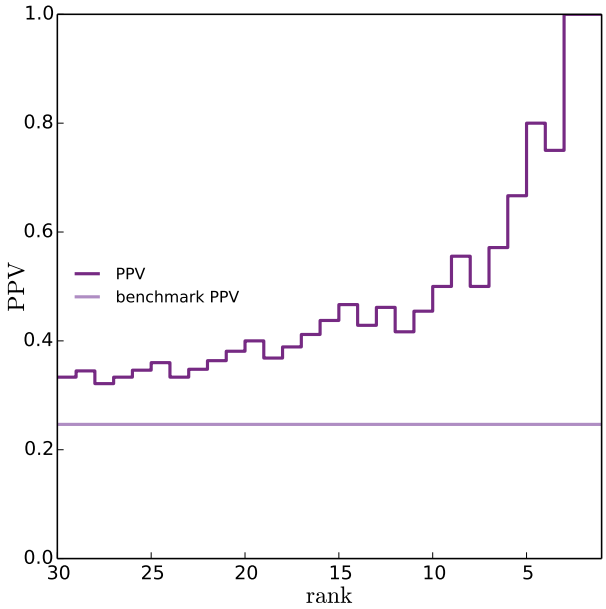

FIG. S7: Performance on the classification problem of identifying pairs of sites where one or more sites is associated with major drug resistance using the top 30 ranked couplings, measured by positive predictive value (PPV).

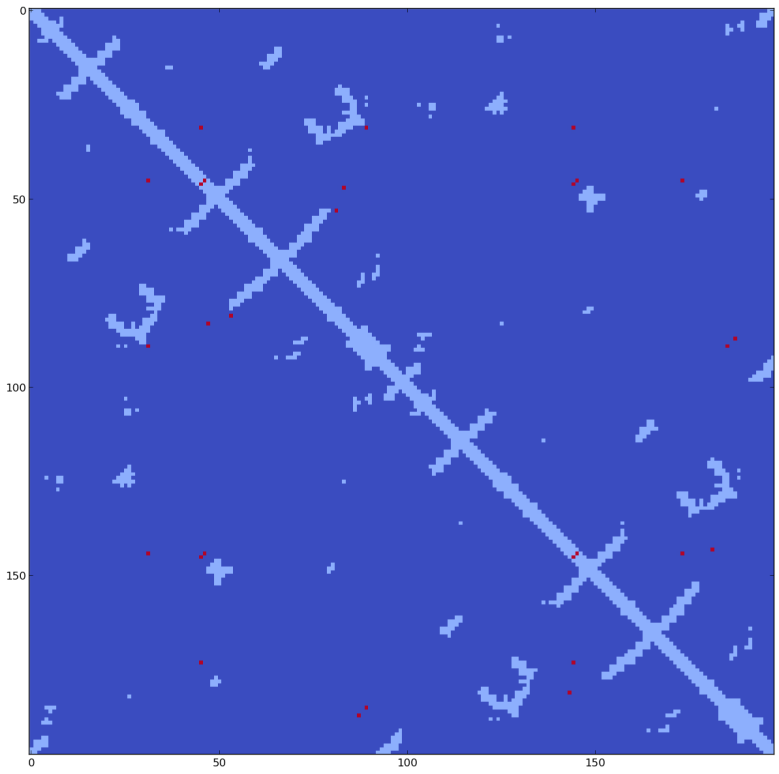

FIG. S8: Contact map for the HIV protease homodimer. Residues in contact are shown in light blue. Top 14 largest couplings are shown in red.

X. PARETO OPTIMAL PROTEASE INHIBITOR PAIRS

Optimal pairs of protease inhibitors are defined as pairs of protease inhibitors that have the least fitness advantageous average couplings in the prevalence landscape, and have the maximum number of distinct, nonoverlapping mutations with the highest levels of phenotypic resistance, as reported in the Stanford Drug Resistance Database [13]. This is to require the maximum number of new mutations for resistance between drugs, and also to ensure that there is as little positive coupling between the sets of resistance mutations as possible.

Here the number of nonoverlapping resistance sites for each pair of protease inhibitors is simply given by the total number of resistance sites for both protease inhibitors together, minus the number of resistance sites that they share in common. The average interaction strength is the average of the coupling strengths (J_{ij}) between all resistance mutations shared by both protease inhibitors. A drug pair is then considered optimal if it cannot improve on one measure of optimality without reduction in another (Pareto optimality). All drug pairs are plotted in Fig. S9. We found 3 optimal pairs: atazanivir-indinavir, atazanavir-fosamprenavir, and darunavir-nelfanavir. Other near-optimal pairs typically include atazanavir, in line with clinical knowledge that the resistance profile of atazanavir tends to be distinct from other protease inhibitors [14].

[1] Leitner T, Korber B, Daniels M, Calef C, Foley B (2005) HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV Sequence Compendium* 2005:41–48.

[2] Barton J, Cocco S (2013) Ising models for neural activity inferred via selective cluster expansion: structural and coding

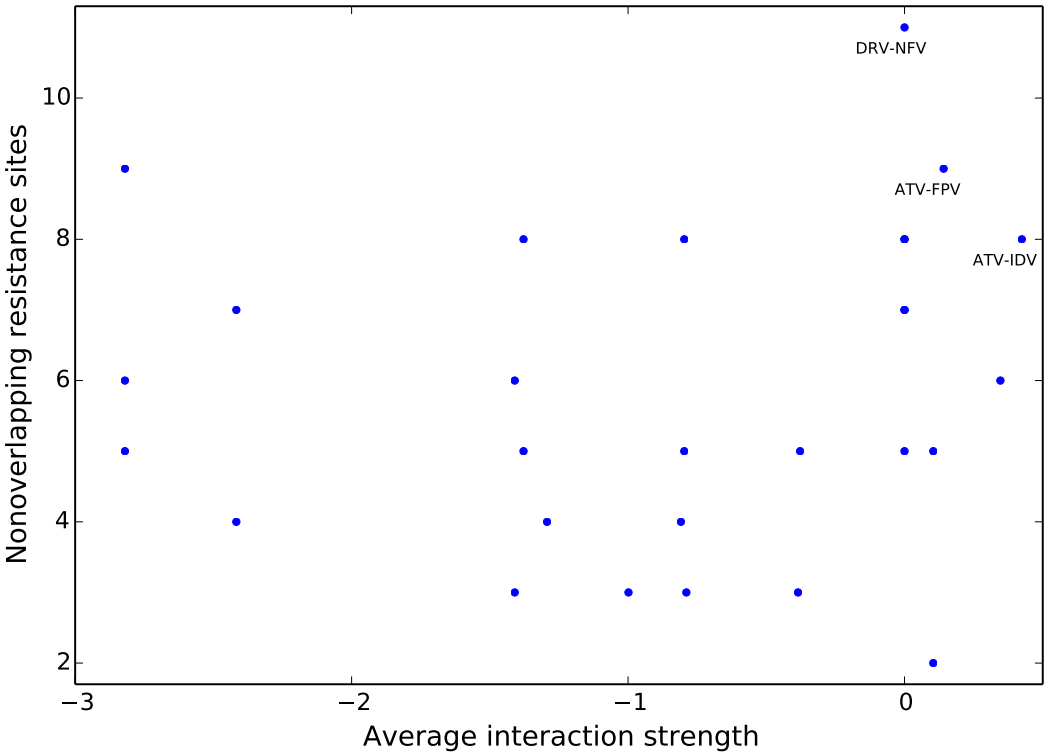

FIG. S9: Pairs of drugs with coordinates given by average energy induced by the presence of resistance mutations to both drugs, and minimum mutual overlap. Pareto optimal pairs are labelled.

properties. *Journal of Statistical Mechanics: Theory and Experiment* 2013:P03002.

[3] Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Physical Review Letters* 106:090601.

[4] Cocco S, Monasson R (2012) Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *Journal of Statistical Physics* 147:252–314.

[5] Leuthäusser I (1986) An exact correspondence between Eigens evolution model and a two-dimensional Ising system. *The Journal of Chemical Physics* 84:1884.

[6] Chang MW, Torbett BE (2011) Accessory mutations maintain stability in drug-resistant HIV-1 protease. *Journal of Molecular Biology* 410:756–760.

[7] Wheeler WH, et al. (2010) Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, US-2006. *AIDS* 24:1203–1212.

[8] Gupta RK, et al. (2012) Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. *The Lancet* 380:1250–1258.

[9] Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* 108:E1293–E1301.

[10] Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nature Biotechnology* 30:1072–1080.

[11] Louis JM, Dyda F, Nashed NT, Kimmel AR, Davies DR (1998) Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry* 37:2105–2110.

[12] Hinkley T, et al. (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics* 43:487–489.

[13] Rhee SY, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* 31:298–303.

[14] Colonno R, et al. (2004) Identification of i50l as the signature atazanavir (atv)-resistance mutation in treatment-naive hiv-1-infected patients receiving atv-containing regimens. *Journal of Infectious Diseases* 189:1802–1810.