Temporal Evolution of the HIV Viral Population under Drug Selection Pressure

Indrani Choudhuri[1,3†], Avik Biswas[2,3†], Allan Haldane[2,3], Ronald M. Levy[1,3*]

[1] Department of Chemistry, Temple University, Philadelphia, PA, USA

[2] Department of Physics, Temple University, Philadelphia, PA, USA

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

1925 N. 12th Street, Philadelphia, PA 19122

[†]Both authors contributed equally

*ronlevy@temple.edu

Abstract

Drug resistance in human immunodeficiency virus (HIV) continues to be a pervasive problem affecting the lives of millions of patients worldwide. In this study, using a Potts protein sequence-covariation statistical energy model combined with Kinetic Monte Carlo (KMC) simulations of evolutionary trajectories under drug selection pressure, we explore the evolution of drug resistance in HIV, starting from an ensemble of drug-naive patient protein sequences. We follow the time course of a total of more than fifty drug resistance mutations in Protease, Reverse Transcriptase, and Integrase. The KMC model on the fitness landscape accurately captures the relative acquisition times of drug-resistance mutations reported in the literature. We find primary drug-resistance mutations with long acquisition times typically require a larger number of accessory mutations to prime the sequence background for their acquisition and eventual entrenchment, while those with shorter acquisition times exhibit pre-existing high initial biases within the drug-naive patient population. We introduce a new metric, the potentiation factor in drug-naive patient protein sequences, that can be an accurate predictor of acquisition times for mutations subject to external selection pressures.

Introduction

The human immunodeficiency virus type 1(HIV-1) currently infects ~40 million people worldwide, and in the absence of a cure, antiretroviral therapy (ART) presents the primary treatment option.[1] However, all antiretroviral drugs, including those from newer drug classes, are at risk of becoming partially or fully inactive due to the emergence of drug-resistance mutations.[2,3,4] The rapid mutation rate of HIV plays a major role in the failure of antiretroviral therapies (ARTs) amongst infected patients, leading to drug-resistance mutations (DRMs) occurring in response to drug selection pressure.[5] Enzymes encoded by the *Pol* gene of HIV, protease (PR), reverse transcriptase (RT), and Integrase (IN) have been the major focus of antiretroviral therapy over the past several decades. The combined effects of the host immune response and selection pressure from ART largely affect the overall fitness landscape of the targeted enzymes (PR, RT and IN), which is modulated by constraints on the proteins' structure, function, thermodynamics, and kinetics.[6,7, 8] As a result, complicated mutation patterns often arise in these proteins located both near and far from the active site[8,9,10] which can be observed in HIV patient protein sequences that are available on large public databases such as the Stanford HIV drug resistance database,[11] and the Los Alamos HIV sequence database.[12] Primary drug-resistance mutations generally occur at a penalty to the fitness of the virus, in the context of the wild-type viral sequence. The effect of a mutation, however, is dependent on the entire genetic background in which it occurs, a phenomenon known as 'epistasis'. Due to epistatic interactions with the sequence background, primary DRMs can become highly favorable in particular sequence backgrounds with the accumulation of patterns of accessory mutations, such that there is a fitness penalty for reversion to the wild-type residue, leading to a kind of evolutionary trapping or "entrenchment" of the primary mutation in its background.[13,14, 15,16] There is also a feedback between "entrenchment" of the primary mutation and its background, with the resultant accumulation of the associated (accessory/compensatory) mutations also dependent on the primary mutation ("contingency") and vice versa. In the presence of drug pressure, this leads to a complex interplay between the functions of the primary and secondary/accessory mutations.[17,18]

Studies have illustrated the effects of epistasis on the fitness landscape of HIV proteins, [14,18,19,20,21] but an understanding of how epistasis causes some DRMs to be acquired rapidly while others are acquired much more slowly, is lacking. In recent work, we introduced a kinetic model (using Kinetic Monte Carlo methods) which describes the evolution of HIV sequences over an epistatic fitness landscape with the help of the *Metropolis* scheme and uses a Potts sequence-covariation statistical energy model[22] inferred from drug-experienced patient protein sequences in the Stanford HIVDB, focusing on the effect of epistasis on sequences in evolutionary equilibrium. In order to establish a baseline for the effect of epistasis on the evolutionary dynamics of HIV, we followed the kinetics of a DRM within the drug-experienced equilibrium ensemble of patient protein sequences.[19] In the current work, we focus on the kinetic routes of emergence of these drug-resistance mutations starting from an ensemble of drug-naive HIV patient sequences, each representative of the consensus sequence in a drug-naive patient, evolving under the influence of drug pressure as captured by the drug-experienced Potts statistical energy model, leading to a much broader distribution in the number of mutations observed in drug-experienced patients (Fig 1).

We focus on ~ 51 major drug-resistance mutations in the three HIV drug-target proteins.[11] Overall, there is a very strong correspondence between predicted acquisition times of DRMs based on the kinetic trajectories and acquisition times reported in the literature. As described in the following section, the evolutionary dynamics of drug-resistance mutations are largely modulated by two factors: Primary DRMs with longer acquisition times generally require many accessory mutations to occur first, priming the sequence background for the acquisition of the DRM. In contrast, most DRMs with short acquisition times are found to exhibit a pre-existing high initial bias for the mutation under the drug-experienced statistical energy model relative to the drug-naive model when averaged over the initial drug-naive patient population. The

Potts model parameterized on drug-experienced HIV patients combined with Kinetic Monte Carlo techniques is a powerful predictor of the evolutionary trajectories leading to drug resistance.

Results

Kinetic Routes to Drug Resistance

In this section we follow the temporal evolution of a series of primary drug resistance mutations in HIV-1 reverse transcriptase (RT), protease (PR), and Integrase (IN) using the kinetic Monte-Carlo method to evolve the sequences with a drug-experienced Potts statistical energy model. Fig 1 shows that the distribution of the total number of mutations per sequence measured with respect to the consensus sequence in reverse transcriptase (RT) evolves from a narrow distribution of sequences with a peak at 6 mutations and a maximum of ~ 20 mutations in the drug-naive patient population to a much broader distribution of sequences with a peak at ~12 mutations and a maximum of more than 30 mutations in the drug-experienced population We use a coarse-grained (See "Computational Methods" section) representation of the within-host evolutionary process which tracks the consensus sequence of a host viral population over time as it accumulates substitutions due to underlying forces. We evaluate the change in the average frequencies of a set drug resistance mutations in RT as the ensemble of sequences shown in Fig. 1 evolves from the drug-naive initial state to the drug-experienced final state (results are also obtained for DRMs in PR and IN). The time to acquire drug resistance is measured as a function of the number of mutations attempted per position (The length of PR, RT and IN are 99, 188 and 263 respectively). The acquisition time is extracted by an exponential fit of the increase in the average DRM frequency starting from the initial drug-naive ensemble and ending in the final drug-experienced state. We found that the temporal evolution of DRMs starting from the drug-naive ensemble is well described by a single exponential. The comparison of the DRM acquisition times extracted from the KMC models with the literature values for RT are presented in. (Fig. 2, Table 1 in the main text and Fig. S2-S3, Table S1-S2 in the supporting information).

We follow $\sim$51 primary drug resistance mutations which have a wide range of acquisition times according to the literature. These include 14 PR DRMs, 13 IN DRMs, and within RT we follow 14 Major Nucleoside RT Inhibitor (NRTI) Resistance Mutations, and 10 Major Non-Nucleoside RT Inhibitor (NNRTI) Resistance Mutations. The RT (NRTI) mutations are listed in table 1 (The IN, RT (NNRTI) and PR mutations are listed in tables S1-S3 respectively). The starting point of the simulation is the drug-naive patient's population, and the equilibrium state represents the drug-experienced patients populations. We track the temporal evolution by following the change in frequency of each primary DRM in PR, RT (NRTI and NNRTI) and IN (total$\sim$51) and calculated the acquisition time (τ) of drug resistance. (Fig. 2, Table 1 in the main text and Fig. S2-S3, Table S1-S2 in the supporting information). The acquisition time is extracted by an exponential fit of the increase in the DRM frequency from initial drug-naive state to the final drug-experienced state, assuming the rise of the frequency is exponential, and the exponential scale parameter gives the timescale.

The analytical results show that most of the Markov systems have exponential approach to equilibrium. We model a single exponential fit as; $Y = \Pi_E - (\Pi_E - \Pi_0)\, e^{(-r/t)}$. This is a simple single exponential model where Π_E =Frequency of the mutation at equilibrium (At drug-experienced state), Π_0 = Initial frequency of the mutation (At drug-naive State) and the time constant (τ) of a DRM is calculated by $\tau = 1/r$. The calculated acquisition time (τ) is then compared with the time in the literature. There is a large literature where a drug or combinations of drugs targeting RT, PR, and IN are given to patients and the sequential appearance of mutations over time are recorded.

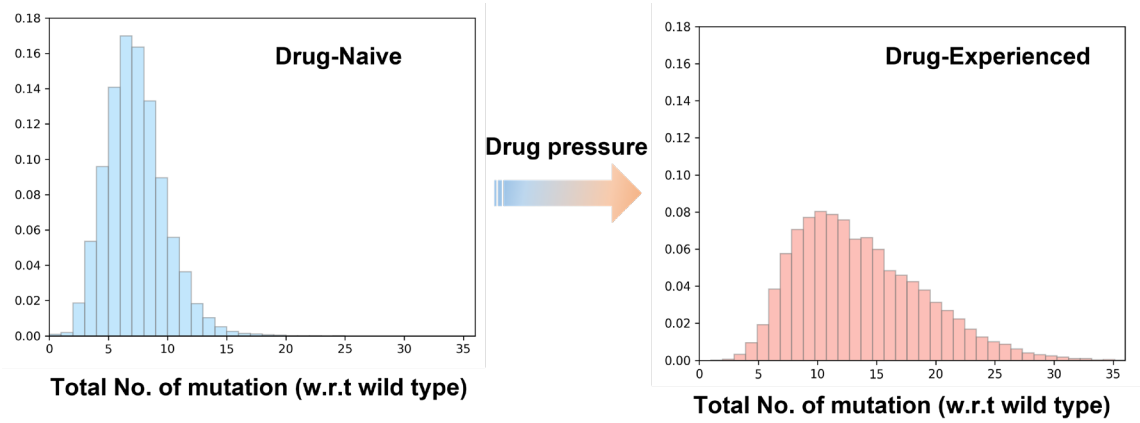

Fig 1. The distribution of the number of inhibitor or drug-associated mutations in HIV-1 Reverse Transcriptase (RT) in (A) the drug-naive multiple sequence alignment (MSA) available from the Stanford HIV database, and (B) the drug-experienced MSA.

We divide the primary DRMs into three categories based on the exponential fits of the number of KMC steps required to reach the mutational frequencies observed in the drug-experienced population; these times are compared with the DRM acquisition times reported in the literature.

Fast: The acquisition time (τ) from our simulation ranges $\sim 1-10$. Mutations emerge in the patients within $\sim 0-3$ months.

Intermediate: The acquisition time (τ) from our simulation ranges $\sim 10-29$. Mutations emerge in the patients within $\sim 3-5$ months.

Slow: The acquisition time (τ) from our simulation ranges $>\sim 30$. Mutations emerge in the patients for more than ~ 6 months.

Based on this timeline classification, we have compared (Table 1) the calculated acquisition time (τ) with the time collected from the literature for 13 NRTIs in RT. Fig. 2 clearly shows the three different classes of DRMs for RT and IN. The range in acquisition times is large; the fastest DRMs are acquired ~20 times more rapidly than the drug resistance mutations which are the slowest to be acquired.

We divide the primary DRMs into three categories based on the exponential fits of the number of KMC steps required to reach the mutational frequencies observed in the drug-experienced population; these times are compared with the DRM acquisition times reported in the literature. For example, (Fig. 2A,C) the K65R mutation which is a major nucleoside RT inhibitor (NRTI) resistance mutations rises very fast ($\tau < 10$) and appears in patients within 2 months.[23] On the other hand, L74I (Fig. 2B,C) (RT/NRTI) rises very slowly ($\tau > 40$) and takes almost a year[24] to appear in patients under drug pressure. We conclude from the results summarized in Fig. 2 and Table 2, that the Kinetic Monte Carlo model of the evolution of drug resistance mutations in RT captures features of the mutational landscape for this HIV-1 enzyme evolving under drug pressure that distinguishes DRMs which are acquired slowly from those acquired rapidly, as reported in the literature.

Similarly, we analyze the kinetics of major drug resistance mutations in different in Protease (PR) and Integrase (IN) (see Table S1 and S3 and Fig. S2-S3). We choose ~14 primary DRMs and drug associates mutations in IN whose acquisition time ranges from a few days to a year. Literature surveys show that, T66I/K[31] E92Q[32] and Y143R[33] mutations in IN appear in the patient population within ~ 3 months of drug therapy. These mutations are categorized in the literature as fast emerging mutations under drug pressure. The KMC model estimates of the calculated time constants (τ) for these mutations ($\tau < 10$) are also fast, consistent the literature reports (Fig. 2).

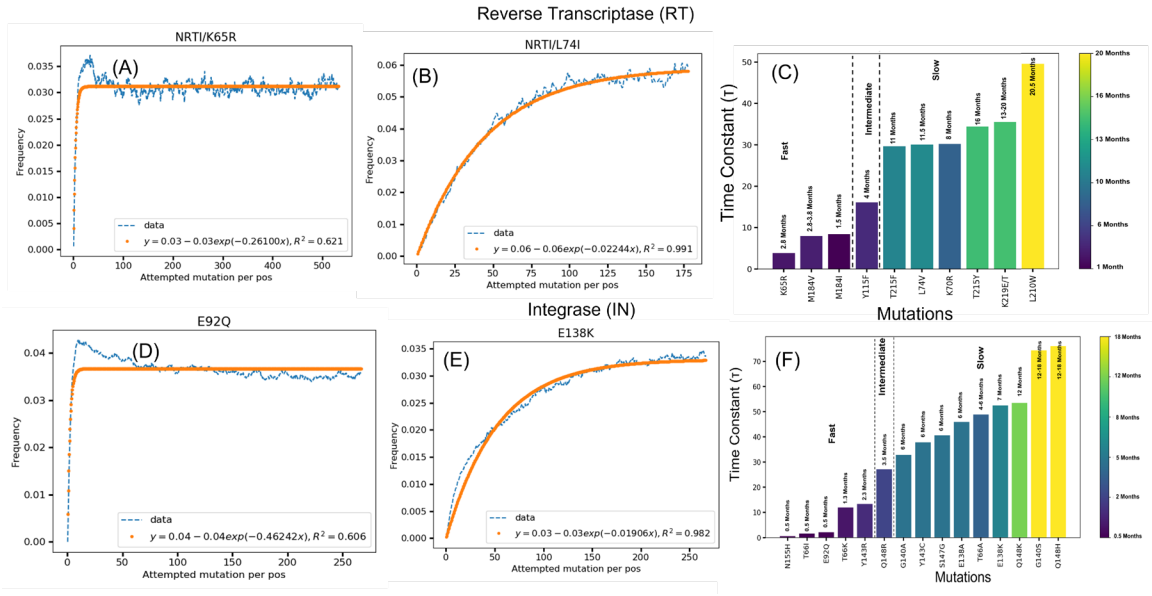

Fig 2. Evolution of drug resistance from drug-naive state to drug-experienced state in major nucleoside RT Inhibitor (NRTI) resistance mutations (upper panel) and Integrase Inhibitor (INSTI) resistance mutations (lower Panel). The change in frequencies of a primary drug resistance mutations (A) K65R, (B) L74I in RT (NRTI) and (D) E92Q, (E) E138K in IN. The simulated is started with drug-naive MSA from Stanford HIV Database[11] and evolved towards drug-experienced state. The evolution time is presented as a function of the number of mutations attempted per position (The length of RT is 188). The literature correspondence of acquisition time (time constant (τ)) of (C) major nucleoside RT Inhibitor (NRTI) resistance mutations and (F)Integrase Inhibitor (INSTI) resistance mutations. The height of each bar represents the time constant (τ) calculated from KMC simulations. The color bar represents the time required for each DRM to emerge in the patient's population under selective drug pressure which is collected from the literature.

On the other hand, E138K[34] Q148K,[35] G140S[36,37] Q148H[33,37] mutations are relatively slow to appear in the patient population as reported in the literature ($\sim 12 - 18$ months) (Fig. 2F). Our study agrees well with the literature as the time constants ($\tau > 30$) for these mutations are very large compared to those for the IN drug resistance mutations acquired rapidly (Fig.2 E-F) and Table S1). The calculated KMC model time constants for drug resistance mutations to appear in PR and for NNRTI mutations induced in RT are also consistent with the times to acquire the specific mutation under selective drug pressure reported in the literature. (Fig.2, S2-S3 and Table S2-S3, Supporting Information). The calculated acquisition time (τ) and the literature timeline of the DRMs are found to be unaffected by the change in the genetic code (Number of nucleotide changes (Table 1)) associated with the certain mutations.

The kinetic Potts model which we used to study the acquisition time (τ) is a fitness based coarse-grained model which represents the within-host evolutionary process (See computational methods). The model is coarse-grained over all drugs to imitate the drug pressure and over the immune pressure of the patients. The most remarkable point is that the kinetic model is successfully able to capture the kinetic notion of a wide range of primary DRMs in different target proteins (PR, RT and IN) in HIV.

Table 1. Acquisition times of emergence of major nucleoside RT inhibitor (NRTI) resistance Mutations using KMC simulation and literature survey

Mutations	Nucleotide Changes	Nature of Mutation	Frequency (%)		Time Constant (τ)	Timeline in Literature (Months)	Classification of Timeline (Literature)	Reference
			Drug-naive	Drug-experienced				
K65R	1	Transition	0.06	3	3.83	~2.8	Fast	23
M184V	1	Transition and Transversion	0.80	54	7.94	~2.7 − 3.8	Fast	25
M184I	1	Transversion	0.05	1	8.40	~1.5	Fast	26
Y115F	1	Transversion	0.03	2	16.05	~4	Intermediate	26
T215F	2	Transition and Transversion	0.17	13	29.64	~11	Slow	27,28
L74V	1	Transversion	0.08	10	30.03	~11.5	Slow	26
K70R	1	Transition	0.30	20	30.19	~8.6	Slow	27
T215Y	2	Transversion	0.36	8	34.37	~15.92	Slow	27
K219E/T	1	Transition	0.18	12	35.51	~11	Slow	27
K219Q	1	Transversion	0.50	32	35.90	~22.5	Slow	27
D67N	1	Transition	0.65	37	43.29	~12	Slow	29,30
L74I	1	Transversion	0.06	6	44.56	~12 − 14	Slow	24
L210W	1	Transversion	0.56	27	49.50	~20.25	Slow	30

Effect of the number of background mutations to acquire resistance

Epistasis in HIV proteins causes strong background-dependent fitness bias or favorability for particular amino acids at each position.[18] This background-dependence is a unique property of the epistatic fitness landscape, and has important implications for evolutionary kinetics over this fitness landscape. In the previous section we classified a wide range of primary DRMs into three classes (Fast, Intermediate and Slow) based on their acquisition time constant (τ) and compared with the literature. Here we explore the effect of mutational background behind the two distinct kinetic behavior of DRMs.

We have parameterized our models using drug-experienced MSAs from the Stanford HIV database, and so the parameterized epistatic interaction strengths reflect this sequence ensemble. Once parameterized, the Potts model provides a log-likelihood function $E(S)$ for each sequence S which is interpreted as a sequence fitness. Critically, this model is "generative", as one can generate new sequences from the model and the mutational covariation statistics of the generated sequences will match those of the HIVDB.[11](See computational methods for more details) The favorability of mutation occurring at a particular position in a given sequence background S can be evaluated by the change in the Potts statistical energy[18,20] of acquiring that specific mutation from the wild-type residue (Eq. 1) at that site.

$$\Delta E(S) = E_{Wildtype}(S) - E_{Mutant}(S) \quad (1)$$

where $\Delta E = 0$ indicates the neutral zone where a specific mutation is neither favored or disfavored by the background. A small range ($-0.5 < \Delta E < 0.5$) near the neutral point ($\Delta E = 0$) is defined as the neutral zone. Sequences whose energy difference fall above the neutral zone ($\Delta E > 0.5$) are entrenching backgrounds favoring the mutation, and whose energy falls below the neutral zone ($\Delta E < -0.5$) are disfavorable, i.e., the mutant sequence fitness is less than the wild type fitness. We emphasize that because of the epistatic nature of the fitness landscape, a mutation will have different favorability in different sequences.

Sequences sampled in the drug-experienced state are particularly interesting due to the abundance of complex mutational patterns which arise in the HIV proteins which are major targets for HIV drug therapy (ART). If $\Delta E > 0.5$, this implies that the mutation is favorable in the given sequence background, and it will lead to fitness penalty associated for reversion back to the wild-type residue. The consistent higher ΔE ($\Delta E >\sim 3.5$) and higher hamming distances (Hamming distance $>\sim 23$) Fig. 3 indicates that the primary mutations leading to drug resistance can become highly favored (or entrenched) by the complex background mutations which is in consistent with one of our earlier work where we have studied a key DRM (L90M) in protease.[18] The pair of DRMs with slow and fast kinetics are selected to study the background dependence behind kinetic behavior in such a way that the frequency of the pair of mutations remains similar in drug-naive and drug-experienced patients populations.

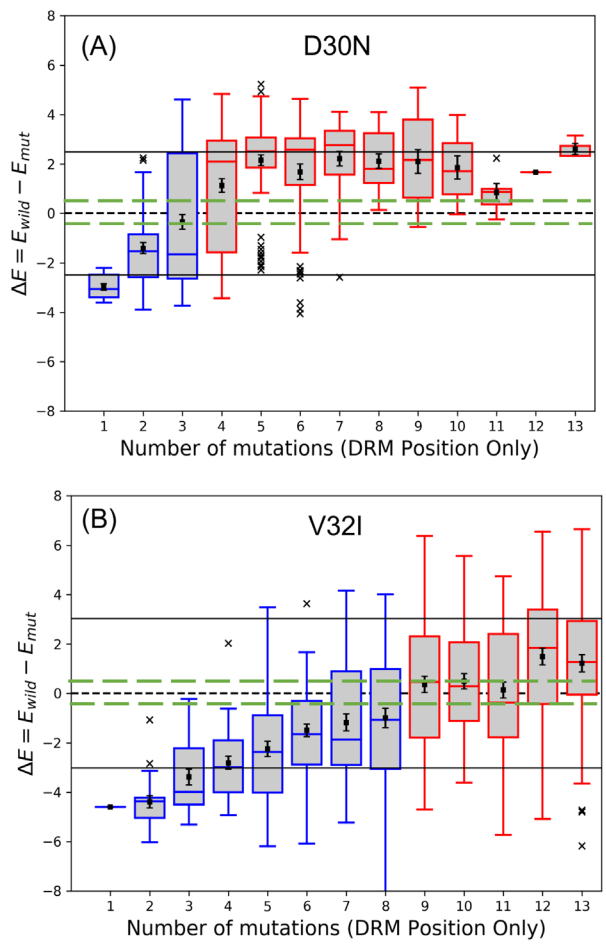

Fig 3. The effect of epistasis on the favorability of a primary resistance mutation in protease (PR). Change of Potts energy difference (ΔE) for a mutation of sequences with the particular mutation, conditional on the total number mutations with respect to wild type consensus considering only the drug resistance positions ($\sim$36) is shown as boxplots annotating the first, second and third interquartile range. The Whiskers extend to 1.5 times the interquartile range with outliers marked as 'x's and the mean values are marked as squares. The left ordinate scale shows the Potts energy difference (ΔE. The area around $\Delta E = 0$ indicated with green dashed lines ($-0.5 < \Delta E < 0.5$) defined the neutral zone. The sequences whose energy difference falls above $\Delta E = 0.5$ (short-dashed line) are entrenching backgrounds favoring the mutation. Sequence backgrounds where the mutation is favored on average are shown in red box plots, the others in blue box plots. (A) The mutation D30N with faster kinetics becomes

favorable on average when there are about more than 4 total number of mutations at the drug resistance positions with respect to wild type consensus. (B) The mutation V32I with slower kinetics becomes favorable on average when there are about more than 9 total number of mutations at the drug resistance positions with respect to wild type consensus. The black dashed line represents $\Delta E = 0$. The black solid lines represent the standard deviation for ΔE.

We begin by reviewing how epistasis causes background dependent amino acid biases at each position in HIV sequences, focusing on effect of epistasis in a pair of HIV DRMs (D30N and V32I) with very different dynamics (Fast and slow) for different target proteins (PR, RT and IN). Among many drug resistance mutation, the effect of epistasis on the favorability are studied for D30N (with "Fast" kinetic) and V32I (with "slow" kinetics) which have similar frequencies in drug-naive and drug-experienced patients populations but with completely different evolution dynamics. D30N and V32I are two primary drug resistance mutation in HIV-1 Protease (PR). We show (Fig. 3) the predicted favorability of D30N and V32I mutations as a function of the number of other mutations,[19, 20] relative to the HIV consensus sequence, for sequences which have these mutations in the drug-experienced MSA collected from Stanford DB. The number of mutations is calculated by considering only drug resistance positions ($\sim$36 positions) in PR.

For the wild-type or close-to wild-type background sequences in the database, both the mutation D30N and V32I are highly disfavored. However, with the help of a small number of accessory mutations ($\sim$4) in the background D30N become increasingly favored (Fig. 3(A)). As D30N required a small number of accessory mutations in background to become favored it boost its dynamic evolution under drug pressure to acquire drug resistance. We closely look at the subset of sequences which have more than 10 mutations considering only $\sim$36 drug resistance positions in the drug-experienced state (Fig. S4, Table S4, Supporting Information). The fraction of sequences within this subset having slower V32I mutations (51.4%) is way above than the faster D30N (4.3%). Provided both the primary DRMs have similar frequency in drug-naive and drug-experienced patients populations. It shows that V32I needs a greater number of accessory mutations in the background, to be increasingly favored, and finally becoming entrenched $\Delta E > 0$ when there are $\sim$9 or more mutations with respect to wild type consensus sequences in HIV-1 (Fig. 3(B)). As accumulation of greater number of accessory mutations in the background compare takes longer time which leads to slower kinetics of V32I compare to D30N.

Similar observations are found in other pair of slow and fast mutations in different protein targets such as RT/NRTI in HIV-1 (Fig. S5, Table S5, Supporting Informations). Where L74I (mutation with slower kinetics) needs to accumulate larger number of accessory mutations in the background compared to K65R (fast mutations) to achieve drug resistance in drug-experienced patients population. We emphasize that Fig. 3 (Fig. S5 in supporting information) does not strictly show a time-ordering of sequences, as this figure shows independent sequences sampled from different patients, rather than a set of sequences sampled from a single patient over time. However, it is commonly observed that a host's HIV population consensus sequence is initially close to wild-type HIV before drug exposure but accumulates additional accessory mutations over time after drug exposure. In other words, the vial consensus sequence in a host typically progresses from sequences like those in the left end in Fig. 3 to those in the right end with more mutations, causing the primary DRM to become favorable or "trapped" when it was not so initially.

Though, it is not just the number of mutations but likelihood of the mutation in a specific background also play an important role to potentiate the dynamic evolution mutation in the drug-naive patients population when it undergoes anti-retrovital therapy (ART). In the next section we will study the role of the likelihood of different DRMs towards their kinetics supported by drug pressure.

Pre-existing bias in drug-naive patient protein sequences to acquire resistance

In this section, we study the impact of the drug-naive sequence background on the kinetics of resistance mutations. Across the diversity of sequence backgrounds observed in drug-naive patients, there exist certain backgrounds that may be highly primed or highly supportive towards acquiring drug-resistance mutations when subjected to drug selection pressure. Under this condition, a resistance mutation can be acquired and eventually trapped or entrenched relatively faster in these highly supportive sequence backgrounds leading to faster kinetics for such mutations. We define this pre-existing initial bias in drug-naive patient protein populations towards acquiring particular drug-resistance mutations as the corresponding potentiation factor, PF, for the mutation. Quantitatively the PF for a mutation M^i at position *i* in sequence *S* is defined as:

$$PF = P_{exper}\,(M^i | S^i_{naive}) \qquad (2)$$

where S^i_{naive} is the set of sequences sampled from drug-naive patients, and the bias for mutation *M* at position *i* in a given sequence background is calculated as $fra_{exp}(-E_{S_i^M} \sum_\beta \exp\,(-E_{S_i^\beta}))$.

In other words, the PF represents the likelihood of a DRM in sequence backgrounds sampled from drug-naive patients with the introduction of drug-selection pressure, which is then reinforced as the sequences evolves under drug pressure.

The scenario can be compared to a biased race-track, where drug-resistance mutations with pre- existing high biases in drug-naive sequence backgrounds, or high PF, closer to their observed frequencies in drug-experienced patient populations, are analogous to runners with a starting point much closer to the finish line than others (Fig. 4A). Under assumptions of similar evolutionary growth rates, these mutations are expected to rise to their drug-experienced frequencies with shorter time-constants under our kinetic simulations. In Fig. 5, we compare the ratio of drug resistance mutation frequencies in drug-experienced HIV patient sequences to the corresponding PF in drug-naive patients, with the respective time constants for evolution under drug-selection pressure, with statistically significant correlation between the two. This illustrates that the closer the initial bias is to the drug-experienced frequencies, the faster is the rise of the DRM (under drug-selection pressure) much like the analogy of a biased race-track, where the runners start at different positions reaching the finish line in accordance with their biased starting positions.

The drug-experienced Potts Hamiltonian is cognizant at of the effective fitness landscape of the virus under an external selection pressure on the evolution of the virus, forcing the virus to explore regions of its intrinsic fitness landscape that are less accessible in order to avoid killing by the drug. While the drug-naive Potts Hamiltonian accurately captures the bias across sequence backgrounds for mutations in the drug-naive state (Supplementary Table S6-S9), it is not predictive of the response of such sequence backgrounds to drug selection pressure, or the time scales of the kinetic evolution of DRMs to their respective drug-experienced frequencies (Supplementary Fig. S7). The drug-experienced Potts Hamiltonian however accurately captures the pre-existing bias for DRMs in the drug-naive patient population and the time scales for their kinetic evolution, and is thus, predictive of their response to drug selection pressure. It is also expected that those sequence backgrounds with high initial biases for resistance mutations are more likely to be observed under drug-pressure, while in those with low initial biases, specific drug-pressure mediated mutation pathways lead to the eventual entrenchment of DRMs with slower kinetics. While beyond the scope of the current study, the likelihood of observing in drug-experienced patient samples, sequence backgrounds which are highly supportive of resistance mutations (high initial bias) under drug

selection pressure in drug-naive patients will be evaluated using canonical reweighting techniques in future studies.

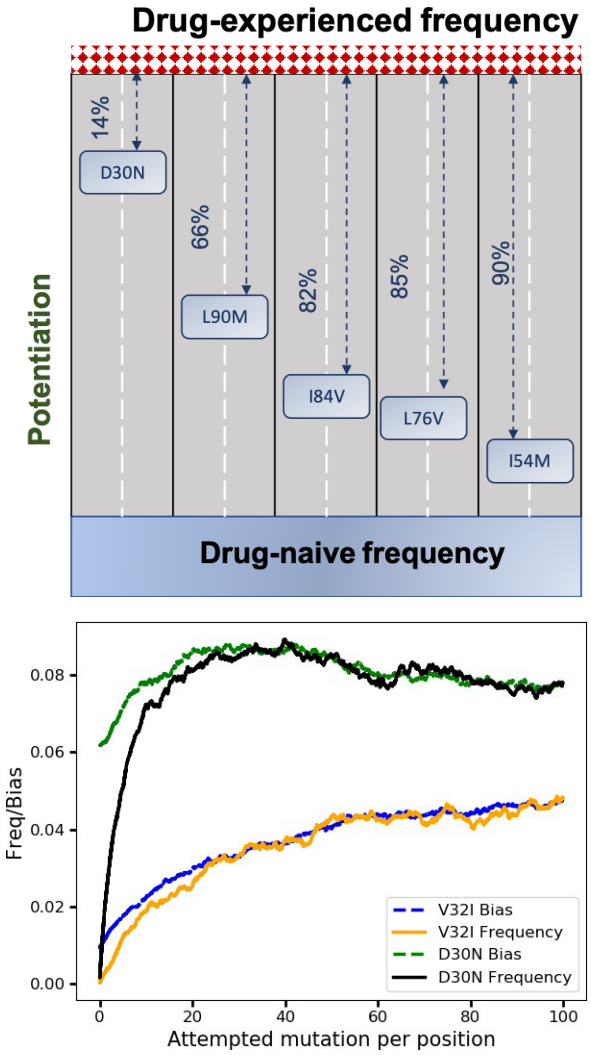

Fig 4. Figure (A) shows the analogy with a biased race-track, where the runners are HIV protease drug-resistance mutations with different starting points relative to the finish line (frequencies in drug-experienced patients), determined by their potentiation factors in drug-naive HIV patient protein sequences. The relative distances to the finish line from their actual drug-naive frequencies are indicated in blue above the mutation labels. DRMs with higher potentiation factors are expected to reach the finish line faster than others. Figure (B) shows the temporal evolution of the bias and frequency of a mutation as a function of the average number of attempted mutations per position for two differently evolving DRMs, D30N (fast), and V32I (slow) in HIV-1 protease. As in the previous analogy, the bias for the faster evolving mutation is much closer to the final drug-experienced frequency, and the frequency quickly evolves to catch up with the bias in the sequence background, whereas for slower evolving mutations, the bias and frequencies are both far from the final and evolve closely together to reach the final drug-experienced frequency.

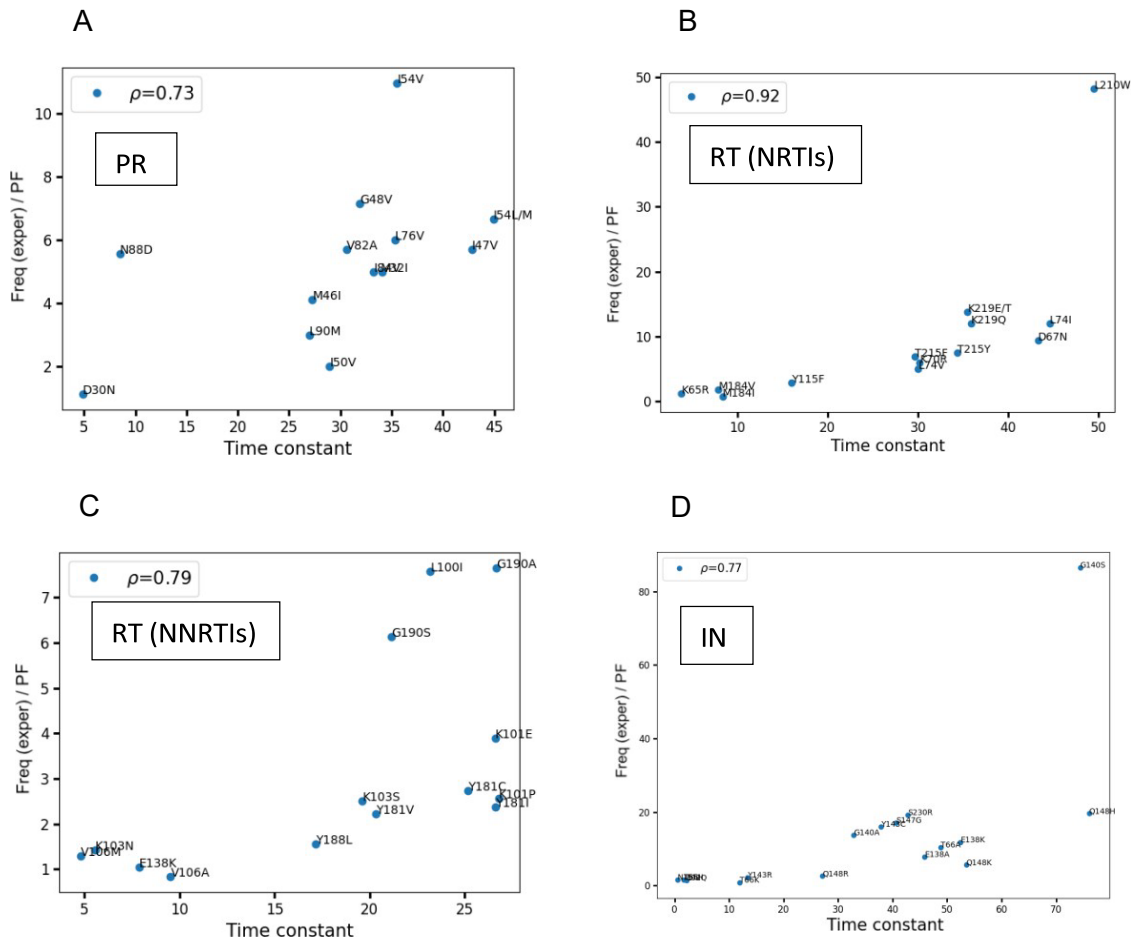

Fig 5. Correlation between the ratio of frequency to PF and time constants (τ) for evolution.

Discussion

The evolution of HIV under immune and drug pressure induces correlated mutations which significantly changes the drug susceptibility of a protein and called drug resistance mutations (DRMs). The kinetic pathways of the emergence of the DRMs under selective drug pressure which can able to alter the fitness landscape of a protein where more rugged region of the landscape otherwise not accessible to the virus earlier, can become accessible as DRMs offer added advantages under therapeutic pressure[5,38,39]some of our previous studies, we focus on the correlated mutations encoded in the MSA of drug-experienced patients population and how the DRMs become highly favored (or trapped) by the complex mutation background in different protein targets[14,18,20,38]

Literature survey shows that the timeline of emergence of DRMs among the drug-naive patients population varied from a few weeks to a year or more when these patients receive antiretroviral therapy (ARTs). We evaluate the kinetic pathways behind the fast and slow emergence of the drug resistance mutations (DRMs) among the drug-naive patients under drug pressure using fitness-landscape based Potts covariation model of epistatic with Kinetic Monte Carlo simulation of evolutionary trajectories. We have selected 51 DRMs from three different protein targets (PR, RT (NRTI and NNRTI), IN) to study their dynamics by calculating the acquisition time (τ) and compared them with the timeline of emergence reported in the literature. based on the timeline reported in the literature we classified the kinetics of the DRMs in three categories such as, “Fast”, “Intermediate” and “Slow”. The DRMs with low acquisition time (τ < 10) falls under the “Fast” category with emergence time ranges 0.5-3 months in the literature.

Whereas, the DRMs are with high acquisition time (τ > 30) are categorized as mutations with slower emergence under drug pressure according to the literature. The reported evolution time of the slow DRMs varied from 8 to 20 months. These classifications are valid for a large number of DRMs in different targeted proteins (PR, RT and IN). It is remarkable to see that our fitness-based course-gained Potts model accurately capture the kinetics of the targeted DRMs in various protein targets in HIV.

It is evidenced from out earlier works[18,20] that epistasis in HIV proteins causes strong background-dependent fitness bias or favorability for particular amino acids at each position. Here, we show that the importance of the background-dependence on the evolutionary kinetics of the DRM mainly relay on the number of mutations in the background. Our study shows that those DRMs with faster kinetic required fewer mutations to be favored in the background. The requirement of larger accessory mutations for a mutation to be favoured by the background force to take a slower kinetic routes or higher acquisition time (τ) to emergence in the patients when then undergoes anitiretroviral therapy.

Across the diversity of sequence backgrounds observed in drug-naive patients, there exist certain backgrounds that may be highly primed or highly supportive towards acquiring drug-resistance mutations when subjected to drug selection pressure. Under this condition, a resistance mutation can be acquired and eventually trapped or entrenched relatively faster in these highly supportive sequence backgrounds leading to faster kinetics for such mutations. we introduce a new metric called as potentiation factor (PF) which is the measure of pre-existing initial bias of a mutation within the drug-naive patients population to understand the origin of different kinetics of different mutations. The scenario can be compared to a biased race-track, where drug-resistance mutations with pre-existing high biases in drug-naive sequence backgrounds, or high PF, closer to their observed frequencies in drug-experienced patient populations, are analogous to runners with a starting point much closer to the finish line than others (Fig. 4A). Under assumptions of similar evolutionary growth rates, these mutations are expected to rise to their drug-experienced frequencies with shorter time-constants under our kinetic simulations. The drug-experienced Potts Hamiltonian accurately captures the pre-existing bias (Table S6-S9, Supporting Information) for DRMs in the drug-naive patient population and the time scales for their kinetic evolution, and is thus, predictive of their response to drug selection pressure. It is also expected that those sequence backgrounds with high initial biases for resistance mutations are more likely to be observed under drug-pressure, while in those with low initial biases, specific drug-pressure mediated mutation pathways lead to the eventual entrenchment of DRMs with slower kinetics. The study is envisioned to comprehend the novel pathways in HIV drug resistance that arise in response to the treatment of clinically used drugs highlighting the primary drug resistance mutations and providing strategies to predict future variants and detection and exploring the scope of stratified medicine.

Computational Methods

In this section, we will discuss about the Potts Hamiltonian model and the motivation behind the model. We also describe the Kinetic Monte Carlo methods which is used for temporal evolution study.[18,20]

Potts Hamiltonian Model

We use the Potts model which is a probabilistic model designed to describe the probabilities of observing specific states of a system that is constructed to be as unbiased as possible except to agree with the average first- and second-order observable (marginals) from the data. In a set of protein sequences, the single and pair amino acid frequencies are average quantities that can be estimated from the finite samples using the data. The details of the models are described in our previous works.

Kinetic Monte Carlo (KMC) simulations

The Kinetic Monte Carlo simulation is a Monte Carlo method which is intended to simulate the time evolution processes with known transition rates between states.

Metropolis algorithm[40] is used to evaluate the metropolis acceptance probability of a mutation such as W (Wild-Type)→M (Mutant) at a randomly chosen position *i* in a given sequence background at every simulation step given by $f_{W\to M}^{MFT}$

$$f_{W\to M}^{MET} = min\{1, P_M/P_W\}$$
$$= min\{1, e^{\Delta E_{W\to M}}\} \quad (3)$$

where $\Delta E_{W\to M} = E_W - E_M$ is the change in Potts energy in going from residue W to M in the given background.

At the beginning of the simulation process a set of seed sequences (Drug-naive sequences) are taken and a random position *i* and random mutant residue *α* (Reduced alphabet of 4 letters) are used are chosen the amino acid character at the chosen position *i* is either preserved or mutated based at the chosen position *i* is either preserved or mutated based on based on the Metropolis acceptance rate $f_{W\to M.}^{MFT}$ For example, a mutation V32I from V (Valine) → I(Isoleucine) at position 32 in HIV-1 Protease (99 residues long) has a probability $\binom{99}{1}e^{\Delta E_{L\to M}^{32}}$ associated with the mutation each KMC step. The algorithm used here, allowed self-mutations during the simulation.

Our kinetic Potts model is as a coarse-grained representation of the within-host evolutionary process. Evolution within a host is driven by various selective forces due to host immune response or to maintain viral viability, non-selective forces such as genetic drift, and is affected by other well-known aspects of HIV infection such as retroviral dormancy, compartmentalization, and high recombination rates. In the chronic phase of HIV infection, a large number of viral variants may be present at any point in time, however this population is typically measured to have high sequence identity of close to 99% between pairs of viral genomes in a single host. This is much larger than the typical sequence identity of consensus sequences from different hosts of 90% for our datasets, and justifies summarizing a host's viral population by a single "consensus" sequence. Additionally, the host consensus sequence is observed to accumulate substitutions in a relatively clock like manner, suggestive of sequential selective sweeps. Therefore, instead of modelling the detailed "microscopic" evolutionary forces we use a coarse-graining which only tracks the consensus sequence of a host viral population over time as it accumulates substitutions due to these underlying forces.

In this way, one interpretation of our coarse-grained kinetics is that it models a series of point- mutation events in a viral population which occur according to a Poissonian mutational process, and these mutations are either fixed or lost from the population according to a fitness landscape inferred based on between-host sequence data. We coarse-grain a number of aspects of evolutionary dynamics, for instance we model amino-acid sequences instead of nucleotide sequences, assuming all amino acids can mutate to all others as is commonly done in phylogenetic analyses, for instance in the WAG and JTT models. While this coarse-grained model is necessarily a simplification of HIV viral dynamics, there are key properties of its construction which support conclusions drawn from it. First, inferred "Potts" model prevalence landscape will implicitly capture many of the averaged effects of various microscopic evolutionary forces, because it is fit to HIV sequence datasets which arose under the microscopic dynamics. For instance, it will capture mutational biases as these causes an increase in the inferred prevalence of the biased amino acids. Second, this model is numerically consistent with the observed between-host sequence variation data: If we use this kinetic model to simulate parallel trajectories (representing evolution in different hosts) and collect the final sequences, then by construction the mutational statistics of the generated sequences (frequencies of amino acids and k-mes of amino acids) match those of the cross-host sequence datasets used to train the fitness model. We use a particular inference technique which we have confirmed gives a generative model which very closely reproduces the natural patterns of HIV sequence variation

for high order k-mes in generated sequences.

We assume an underlying Poisson mutational process, such that mutation arises at a rate μ. We implemented this by assigning each step in the Metropolis Algorithm a time drawn from an exponential distribution which is the waiting time for a Poisson process.

$$P(t \mid \mu) = \mu e^{-\mu t} \tag{4}$$

With this overall KMC scheme out simulations match empirical data in two ways. First, the value of μ can be calibrated so that the simulated evolutionary trajectories accumulated substitutions at the same rate as observed experimentally. Second, when using this scheme to run many parallel trajectories until equilibrium, the bivariate residues of the resulting MSA match those observed in the HIV sequence database used to train the Potts model.

Supporting Information

The supporting Information is available.

Acknowledgement

This work has been supported by the National Institutes of Health through grants awarded to Ronald M. Levy (U54-AI150472 subcontract and U54-AI170855 subcontract, R35-GM132090, S10OD020095). The National Science Foundation also provided funding through a grant awarded to RML and AH (1934848). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

(1) UNAIDS. *UNAIDS*. http://www.unaids.org/ (accessed 2022 September 15).

(2) Coffin, J. M. HIV Population Dynamics in Vivo: Implications for Genetic Variation, Pathogenesis, and Therapy. *Science* **1995**, *267* (5197), 483-489. DOI: 10.1126/science.7824947 (acccessed 2022/09/20).

(3) Perelson, A. S.; Neumann, A. U.; Markowitz, M.; Leonard, J. M.; Ho, D. D. HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time. *Science* **1996**, *271* (5255), 1582-1586. DOI: 10.1126/science.271.5255.1582 (acccessed 2022/09/20).

(4) Rhee, S.-Y.; Kassaye, S. G.; Jordan, M. R.; Kouamou, V.; Katzenstein, D.; Shafer, R. W. Public availability of HIV-1 drug resistance sequence and treatment data: a systematic review. *The Lancet Microbe* **2022**, *3* (5), e392-e398. DOI: 10.1016/S2666-5247(21)00250-0 (acccessed 2022/09/20).

(5) Cuevas, J. M.; Geller, R.; Garijo, R.; López-Aldeguer, J.; Sanjuán, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLOS Biology* **2015**, *13* (9), e1002251. DOI: 10.1371/journal.pbio.1002251.

(6) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286* (5438), 295-299. DOI: 10.1126/science.286.5438.295 (acccessed 2022/09/20).

(7) Bloom, J. D.; Gong, L. I.; Baltimore, D. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science* **2010**, *328* (5983), 1272-1275. DOI: 10.1126/science.1187816.

(8) Haq, O.; Andrec, M.; Morozov, A. V.; Levy, R. M. Correlated Electrostatic Mutations Provide a Reservoir of Stability in HIV Protease. *PLoS Computational Biology* **2012**, *8* (9), e1002675. DOI: 10.1371/journal.pcbi.1002675.

(9) Chang, M. W.; Torbett, B. E. Accessory Mutations Maintain Stability in Drug-Resistant HIV-1 Protease. *Journal of Molecular Biology* **2011**, *410* (4), 756-760. DOI: 10.1016/j.jmb.2011.03.038.

(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 Patient Isolates Reveals Patterns of Correlated Mutations in Gag and Protease. *PLOS Computational Biology* **2015**, *11* (4), e1004249. DOI: 10.1371/journal.pcbi.1004249.

(11) Stnaford. *Stanford HIV Database*. https://hivdb.stanford.edu/ (accessed September 15, 2022).

(12) database, L. A. H. s. *Los Alamos HIV sequence database*. http://www.hiv.lanl.gov/ (accessed 2022 September 15).

(13) Pollock, D. D.; Thiltgen, G.; Goldstein, R. A. Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences* **2012**, *109* (21), E1352-E1359. DOI: 10.1073/pnas.1120084109.

(14) Shah, P.; McCandlish, D. M.; Plotkin, J. B. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences* **2015**, *112* (25), E3226-E3235. DOI: 10.1073/pnas.1412933112.

(15) McCandlish, D. M.; Shah, P.; Plotkin, J. B. Epistasis and the Dynamics of Reversion in Molecular Evolution. *Genetics* **2016**, *203* (3), 1335-1351. DOI: 10.1534/genetics.116.188961.

(16) Depristo, M. A.; Weinreich, D. M.; Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics* **2005**, *6* (9), 678-687. DOI: 10.1038/nrg1672.

(17) Yilmaz, N. K.; Schiffer, C. A. Drug Resistance to HIV-1 Protease Inhibitors: Molecular Mechanisms and Substrate Coevolution. Springer International Publishing, 2017; pp 535-544.

(18) Biswas, A.; Haldane, A.; Arnold, E.; Levy, R. M. Epistasis and entrenchment of drug resistance in HIV-1 subtype B. *eLife* **2019**, *8*. DOI: 10.7554/elife.50524.

(19) Choudhuri, I.; Biswas, A.; Haldane, A.; Levy, R. M. Contingency and Entrenchment of drug-resistance mutations in HIV viral proteins. 2022.[In Press, J. Phys. Chem. B]

(20) Flynn, W. F.; Haldane, A.; Torbett, B. E.; Levy, R. M. Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease. *Molecular Biology and Evolution* **2017**, *34* (6), 1291-1306. DOI: 10.1093/molbev/msx095.

(21) Biswas, A.; Haldane, A.; Levy, R. M. Limits to detecting epistasis in the fitness landscape of HIV. *PLOS ONE* **2022**, *17* (1), e0262314. DOI: 10.1371/journal.pone.0262314.

(22) Haldane, A.; Levy, R. M. Mi3-GPU: MCMC-based inverse Ising inference on GPUs for protein covariation analysis. *Computer Physics Communications* **2021**, *260*, 107312. DOI: https://doi.org/10.1016/j.cpc.2020.107312.

(23) GarcíA-Lerma, J. G.; Macinnes, H.; Bennett, D.; Reid, P.; Nidtha, S.; Weinstock, H.; Kaplan, J. E.; Heneine, W. A Novel Genetic Pathway of Human Immunodeficiency Virus Type 1 Resistance to Stavudine Mediated by the K65R Mutation. *Journal of Virology* **2003**, *77* (10), 5685-5693. DOI: 10.1128/jvi.77.10.5685-5693.2003.

(24) Wirden, M.; Lambert-Niclot, S.; Marcelin, A.-G.; Schneider, L.; Ait-Mohand, H.; Brunet, C.; Angleraud, F.; Amard, S.; Katlama, C.; Calvez, V. Antiretroviral combinations implicated in emergence of the L74I and L74V resistance mutations in HIV-1-infected patients. *AIDS* **2009**, *23* (1).

(25) Tisdale, M.; Alnadaf, T.; Cousens, D. Combination of mutations in human immunodeficiency virus type 1 reverse transcriptase required for resistance to the carbocyclic nucleoside 1592U89. *Antimicrobial Agents and Chemotherapy* **1997**, *41* (5), 1094-1098. DOI: 10.1128/AAC.41.5.1094 (acccessed 2022/09/20).

(26) Margot, N. A.; Waters, J. M.; Miller, M. D. In Vitro Human Immunodeficiency Virus Type 1 Resistance Selections with Combinations of Tenofovir and Emtricitabine or Abacavir and Lamivudine. *Antimicrobial Agents and Chemotherapy* **2006**, *50* (12), 4087-4095. DOI: 10.1128/aac.00816-06.

(27) Hooker, D. J.; Tachedjian, G.; Solomon, A. E.; Gurusinghe, A. D.; Land, S.; Birch, C.; Anderson, J. L.; Roy, B. M.; Arnold, E.; Deacon, N. J. An in vivo mutation from leucine to tryptophan at position 210 in human immunodeficiency virus type 1 reverse transcriptase contributes to high-level resistance to 3'-azido-3'-deoxythymidine. *Journal of Virology* **1996**, *70* (11), 8010-8018. DOI: 10.1128/jvi.70.11.8010-8018.1996.

(28) Yerly, S.; Rakik, A.; Kinloch De Loes, S.; Hirschel, B.; Descamps, D.; Brun-Vézinet, F.; Perrin, L. Switch to Unusual Amino Acids at Codon 215 of the Human Immunodeficiency Virus Type 1 Reverse Transcriptase Gene in Seroconvertors Infected with Zidovudine-Resistant Variants. *Journal of Virology* **1998**, *72* (5), 3520-3523. DOI: 10.1128/JVI.72.5.3520-3523.1998 (acccessed 2022/09/20).

(29) Miller, M. D.; Margot, N.; Lu, B.; Zhong, L.; Chen, S.-S.; Cheng, A.; Wulfsohn, M. Genotypic and Phenotypic Predictors of the Magnitude of Response to Tenofovir Disoproxil Fumarate Treatment in Antiretroviral-Experienced Patients. *The Journal of Infectious Diseases* **2004**, *189* (5), 837-846. DOI: 10.1086/381784 (acccessed 9/20/2022).

(30) Yahi, N.; Tamalet, C.; Tourrès, C.; Tivoli, N.; Fantini, J. Mutation L210W of HIV-1 reverse transcriptase in patients receiving combination therapy. *Journal of Biomedical Science* **2000**, *7* (6), 507-513. DOI: 10.1007/BF02253366.

(31) Goethals, O.; Clayton, R.; Van Ginderen, M.; Vereycken, I.; Wagemans, E.; Geluykens, P.; Dockx, K.; Strijbos, R.; Smits, V.; Vos, A.; et al. Resistance Mutations in Human Immunodeficiency Virus Type 1 Integrase Selected with Elvitegravir Confer Reduced Susceptibility to a Wide Range of Integrase Inhibitors. *Journal of Virology* **2008**, *82* (21), 10366-10374. DOI: 10.1128/jvi.00470-08. Margot, N. A.; Hluhanich, R. M.; Jones, G. S.; Andreatta, K. N.; Tsiang, M.; McColl, D. J.; White, K. L.; Miller, M. D. In vitro resistance selections using elvitegravir, raltegravir, and two metabolites of elvitegravir M1 and M4. *Antiviral Research* **2012**, *93* (2), 288-296. DOI: https://doi.org/10.1016/j.antiviral.2011.12.008.

(32) Winters, M. A.; Lloyd, R. M.; Shafer, R. W.; Kozal, M. J.; Miller, M. D.; Holodniy, M. Development of Elvitegravir Resistance and Linkage of Integrase Inhibitor Mutations with Protease and Reverse Transcriptase Resistance Mutations. *PLoS ONE* **2012**, *7* (7), e40514. DOI: 10.1371/journal.pone.0040514.

(33) Fransen, S.; Gupta, S.; Frantzell, A.; Petropoulos, C. J.; Huang, W. Substitutions at Amino Acid Positions 143, 148, and 155 of HIV-1 Integrase Define Distinct Genetic Barriers to Raltegravir Resistance In Vivo. *Journal of Virology* **2012**, *86* (13), 7249-7255. DOI: 10.1128/jvi.06618-11.

(34) Naeger, L. K.; Harrington, P.; Komatsu, T.; Deming, D. Effect of Dolutegravir Functional Monotherapy on HIV-1 Virological Response in Integrase Strand Transfer Inhibitor Resistant Patients. *Antiviral Therapy* **2015**, *21* (6), 481-488. DOI: 10.3851/IMP3033 (acccessed 2022/09/20).

(35) Cooper, D. A.; Steigbigel, R. T.; Gatell, J. M.; Rockstroh, J. K.; Katlama, C.; Yeni, P.; Lazzarin, A.; Clotet, B.; Kumar, P. N.; Eron, J. E.; et al. Subgroup and Resistance Analyses of Raltegravir for Resistant HIV-1 Infection. *New England Journal of Medicine* **2008**, *359* (4), 355-365. DOI: 10.1056/nejmoa0708978.

(36) Blanco, J.-L.; Varghese, V.; Rhee, S.-Y.; Gatell, J. M.; Shafer, R. W. HIV-1 Integrase Inhibitor Resistance and Its Clinical Implications. *The Journal of Infectious Diseases* **2011**, *203* (9), 1204-1214. DOI: 10.1093/infdis/jir025

(acccessed 9/20/2022).

(37) Parczewski, M.; Bander, D.; Urbańska, A.; Boroń-Kaczmarska, A. HIV-1 integrase resistance among antiretroviral treatment naive and experienced patients from Northwestern Poland. *BMC Infectious Diseases* **2012**, *12* (1), 368. DOI: 10.1186/1471-2334-12-368.

(38) Shafer, R. W.; Schapiro, J. M. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev* **2008**, *10* (2), 67-84. From NLM.

(39) Wensing, A. M.; Calvez, V.; Ceccherini-Silberstein, F.; Charpentier, C.; Günthard, H. F.; Paredes, R.; Shafer, R. W.; Richman, D. D. 2019 update of the drug resistance mutations in HIV-1. *Top Antivir Med* **2019**, *27* (3), 111-121. From NLM.

(40) Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57* (1), 97. DOI: 10.2307/2334940.