THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Contingency and Entrenchment of Drug-Resistance Mutations in HIV Viral Proteins

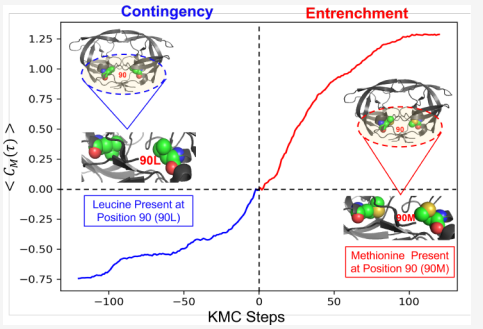
Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

Indrani Choudhuri,[‖] Avik Biswas,[‖] Allan Haldane,[*] and Ronald M. Levy[*]

Cite This: *J. Phys. Chem. B* 2022, 126, 10622−10636

Read Online

ABSTRACT: The ability of HIV-1 to rapidly mutate leads to antiretroviral therapy (ART) failure among infected patients. Drug-resistance mutations (DRMs), which cause a fitness penalty to intrinsic viral fitness, are compensated by accessory mutations with favorable epistatic interactions which cause an evolutionary trapping effect, but the kinetics of this overall process has not been well characterized. Here, using a Potts Hamiltonian model describing epistasis combined with kinetic Monte Carlo simulations of evolutionary trajectories, we explore how epistasis modulates the evolutionary dynamics of HIV DRMs. We show how the occurrence of a drug-resistance mutation is contingent on favorable epistatic interactions with many other residues of the sequence background and that subsequent mutations entrench DRMs. We measure the time-autocorrelation of fluctuations in the likelihood of DRMs due to epistatic coupling with the sequence background, which reveals the presence of two evolutionary processes controlling DRM kinetics with two distinct time scales. Further analysis of waiting times for the evolutionary trapping effect to reverse reveals that the sequences which entrench (trap) a DRM are responsible for the slower time scale. We also quantify the overall strength of epistatic effects on the evolutionary kinetics for different mutations and show these are much larger for DRM positions than polymorphic positions, and we also show that trapping of a DRM is often caused by the collective effect of many accessory mutations, rather than a few strongly coupled ones, suggesting the importance of multiresidue sequence variations in HIV evolution. The analysis presented here provides a framework to explore the kinetic pathways through which viral proteins like HIV evolve under drug-selection pressure.

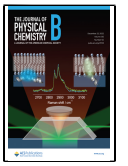


1. INTRODUCTION

HIV evolves rapidly, with studies indicating that in the absence of drug pressure, HIV explores the majority of all single-point mutations for a specific protein within a single patient many times daily,[1−3] rendering the development of an effective vaccine as a significant challenge and driving the emergence of drug-resistant strains through the course of antiretroviral therapy (ART).[4] Recent studies indicate drug resistance develops in up to 68% of patients undergoing monotherapy[5] and in up to 21% of patients undergoing current combination antiretroviral therapy (c-ART),[6] making tackling drug resistance one of the crucial issues in the development of future therapeutics. Although the evolution of HIV, like any other virus, is largely limited by constraints due to function, structural viability, thermodynamics, and kinetics,[7−10] the application of drug selection pressure can force the virus to explore otherwise unfavorable regions of the fitness landscape resulting in complex mutation patterns that arise at residues located both near and far from the drug active site[9,11,12] and provide escape pathways to the virus with a complex interplay in the functions of primary and secondary mutations.[13,14] A detailed understanding of the evolution of these complex patterns of resistance mutations within the drug-experienced HIV-1 population is therefore key to overcoming the issue of drug resistance.

While most drug-resistance mutations (DRMs) naturally occur within the patient viral reservoir in the absence of drug pressure,[15−18] these mutations are unlikely to persist in the drug-naive population, being disfavored in the wild-type viral sequence background(s) due to their detrimental effects on fitness. In the presence of specific patterns of accessory mutations, however, primary drug-resistance mutations can become highly favored by their respective sequence backgrounds with a fitness penalty for reversion back to the wild-type, leading to an evolutionary trapping or "entrenchment" of the mutation.[14,19−23] The entrenchment of drug-resistance mutations is a consequence of, and is modulated by, epistatic

Received: August 26, 2022
Revised: November 18, 2022
Published: December 9, 2022



https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

interactions with the entire sequence background that can strongly influence the evolutionary kinetics of HIV.

Although there have been many studies showing how epistasis shapes the fitness landscapes of viral proteins, including those for HIV,[16,17,24−27] it remains remains poorly characterized how epistatic interactions with the sequence background modulate the kinetics of drug-resistance mutations. Previous work has shown that epistatic interactions in HIV proteins can cause strong background-dependent fitness biases for particular residue types or mutations at each position.[14,19] This background dependence is a unique property of the epistatic fitness landscape and has important implications for evolutionary kinetics over the landscape, but these previous studies have not explicitly modeled the time dependence of the background-dependent biases. Here, we model a sequence of events in discrete time using kinetic Monte Carlo (KMC) simulation. The time is represented in terms of KMC steps or sequence of events (attempted mutation/position) throughout the manuscript.

In this study, we introduce a model of evolutionary kinetics using a Metropolis scheme, to describe evolution over an epistatic fitness landscape we have inferred from observed patient viral sequence data using "Potts" inference, which determines all residue−residue epistatic couplings based on deconvolution of observed mutational covariation. Building on previous work,[14,19] we determine how a mutant coevolves with the sequence background, studying how the sequence background changes in the lead up to the mutation occurring, gradually becoming more favorable on average before it occurs ("contingency"), and how it changes after the mutation occurs, evolving to trap or entrench the mutation ("entrenchment"). We confirm and demonstrate that these contingency and entrenchment effects arise as a consequence of epistasis during evolution over the epistatic HIV fitness landscape inferred from sequence data, by performing ensemble averages over simulated trajectories in "evolutionary equilibrium" over this landscape. While an ultimate practical application of these kinetic models can be to predict mutational pathways in specific sequences in response to drug and immune/vaccine pressure, in this work we limit our focus to the effects of epistasis in "equilibrium" ensemble averages of trajectories, in order to establish baselines for the effects of epistasis on evolutionary kinetics, in order to be able to distinguish these from nonequilibrium behaviors. Fitness is defined relative to an environment, and we focus on viral evolution in the presence of drugs in order to model the behavior of the DRMs (tabulated in Table S1). The aim of this study is to provide a framework to explore the kinetic behavior of the drug-resistance mutations in HIV under a drug treatment environment.

Using ensembles of evolutionary trajectories, we demonstrate the effects of epistatic coupling on kinetics using different metrics. Using the time-autocorrelations of the fluctuations in background-dependent epistatic biases for DRMs we reveal the presence of two evolutionary processes (at a coarse-grained level) that modulate the kinetics of the DRMs with two different time scales. We further support this picture by analysis of the waiting times for the background-dependent mutation biases to revert, demonstrating that the kinetics over the epistatic HIV fitness landscape at shorter time scales are controlled by sequences where the background-dependent biases do not favor or trap the DRM, while for sequences which trap or entrench the DRM, relaxation of the sequence background is necessitated in order to revert the evolutionary trapping effect, resulting in slower kinetics over longer time scales. We also quantify the overall strength of epistatic coupling between residues for different mutations using dispersion indices, showing the importance of epistasis in regulating the waiting time distributions for DRMs to decay back to the most probable residue (wild-type).

We mainly focus on the enzyme Protease (PR), which has been a prominent target of antiretroviral therapies.[28−30] HIV-1 PR is a dimeric enzyme from the family of aspartic proteases[28,29] responsible for processing of the gag and gag-pol polyproteins during virion maturation. The activity of this enzyme is essential for virus infectivity, rendering the protein a major therapeutic target for AIDS treatment.[31,32] We have also extended our study to Reverse Trascriptase (RT) which is an essential enzyme to convert its RNA to viral DNA.

2. COMPUTATIONAL METHODS

In this section, we discuss the Potts Hamiltonian model used in this study. We will also describe the Kinetic Monte Carlo methods used for simulations. The details of waiting time distribution and autocorrelation are discussed in this section.

2.1. Potts Hamiltonian Model. The Potts model is a probabilistic model which aims 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 Potts model method improves on older methods by deconvoluting the "Direct" from "Indirect" mutational interactions, as it is a global model of the observed mutational covariations in multiple sequence alignments (MSAs) of proteins from a common protein family, which arise during the course of evolution through compensatory effects.[33−38] These predicted interactions have been found to correspond well to physical contacts within the 3D structure of proteins, and models inferred from protein sequence data have shown great promise for elucidating the relationship between protein sequence, structure and function.[39−47]

The model is based on the approximation of the unknown empirical probability distribution $P(S)$ which best describes HIV-1 sequences S of length L, where each residue is encoded in an alphabet Q, by a model probability distribution $P^m(S)$.[48] The "least biased" or maximum entropy distribution is considered as the model distribution.[45,49] The maximum entropy model takes the exponential distribution form given below:

$$E(S) = \sum_{i}^{L} h^{i}_{S_i} + \sum_{i<j}^{L(L-1)/2} j^{ij}_{S_i S_j} \tag{1}$$

$$P^m(S) = \exp(-E(S)) \tag{2}$$

where $E(S)$ is the Potts Hamiltonian giving the statistical energy of a sequence S of length L, the model parameters $h^i_{S_i}$ called "fields" represent the statistical energy of residue S_i at position i in sequence S and $j^{ij}_{S_i S_j}$ are "couplings" representing the energy contribution of a position pair i,j. In this form, the Potts Hamiltonian consists of LQ field parameters $h^i_{S_i}$ and $\left(\frac{L}{2}\right)Q^2$ coupling parameters $j^{ij}_{S_i S_j}$ and for the exponential

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

distribution $P^m(S) = \exp(-E(S))$, negative fields and couplings signify favored amino acids. The change in Potts energy due to mutating a residue at position i in S to β is then presented as

$$\Delta E(S^{i}_{\alpha\to\beta)}) = E(S^i_\alpha) - E(S^i_\beta) = h^i_\alpha - h^i_\beta + \sum_{j\neq i}^{L} J^{ij}_{\alpha S_j} - J^{ij}_{\beta S_j} \tag{3}$$

In this form, $(\Delta E(S^i_{\alpha\to\beta}) > 0)$ implies that the residue β is more favorable than residue α at position$i$ for the given sequence S. If α represents the wild-type residue at i and β the mutant, then the mutant is favorable over the wild-type if $\Delta E > 0$ for the change and vice versa. The approach followed in this study is like the one followed in Flynn et al.[19] for HIV-1 PR (for further details on derivation and description of the model parameters see their "Materials and Methods" section as well as their Supporting Information). Our previous study[50] confirms that the models are fit using sufficient data with minimal overfitting.

There are multiple inference algorithms to solve the parameters $J^{ij}_{\alpha\beta}$ which give a model which correctly regulates the input bivariate frequencies. We use the Mi3-GPU software package software to infer the model,[51] as it makes a few approximations and gives a model which accurately recapitulates sequence statistics.

Currently, new machine learning methods are being investigated to analyze large data sets of protein sequences, much like the Potts model. However, we have found that the Potts model accurately captures pairwise and higher order correlations between sites in protein families, better than state-of-the art deep neural networks;[44] this combined with the physical interpretability of the model parameters has led to the use of these models in protein biophysics and evolutionary dynamics. We use the Potts Hamiltonian model for these reasons in this work for studying the evolutionary dynamics of HIV drug-resistance mutations.

2.2. Data Collection and Processing. The protein sequences used in this study are collected from the Stanford HIV database.[52] The filtering criteria used here are HIV-1, subtype B and nonCRF, and drug-experienced (of PI = 1−9 for PR, of NRTI = 1−9), removal of mixtures, and unambiguous amino acid sequences (amino acids are '−ACDEFGHIKLMNPQRSTVWY'). Sequences with insertions ('') and deletions ("∼") are removed. MSA columns and rows with more than 1% gaps ('.') are removed. This resulted in a final MSA size of $N = 5710$ sequences of length $L = 99$ for PR, $N = 19194$ sequences of length $L = 188$ for RT. The detailed discussion of the data processing is available in our previous work.[14]

We obtain the subtype B consensus sequence from the Los Alamos HIV sequence database[53] and consensus and ancestral sequence alignments.[54] The subtype B consensus sequence, which is derived from an alignment of subtype B sequences maintained at the Los Alamos HIV Sequence Database[55] and is a commonly used reference sequence to which new sequences are compared, is used as our reference wild-type (WT) consensus sequence. In the presence of drug pressure, any mutations that occur in patients and affects in vitro drug susceptibility are known as drug-resistance mutations and commonly found in persons experiencing virological failure.[56,57]

2.3. Kinetic Monte Carlo (KMC) Simulations. The KMC technique is a Monte Carlo method intended to simulate the time evolution of processes occurring, typically with known transition rates between states.

Here we have used Metropolis algorithm[58,59] 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 $f^{\mathrm{MET}}_{\mathrm{W}\to\mathrm{M}}$

$$\begin{aligned} f^{\mathrm{MET}}_{\mathrm{W}\to\mathrm{M}} &= \min\{1, P_{\mathrm{M}}/P_{\mathrm{W}}\} \\ &= \min\{1, e^{\Delta}E_{\mathrm{W}\to\mathrm{M}}\} \end{aligned} \tag{4}$$

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

We begin the simulation process with a set of seed sequences, which are Drug experienced sequences. In each step, we choose a random position i (out of L positions) and random residue α (of the four letters in the reduced alphabet, including the current letter at that position). The amino acid character at the chosen position i is either preserved or mutated based on the Metropolis acceptance rate $f^{\mathrm{MET}}_{\alpha\to\beta}$.

HIV sequence data sets are highly conserved, and it is very rare to observe more than 4 different residue types at any position in the HIV proteins. Most amino acids are never seen at each position. Therefore, a reduced alphabet of 4 letters is used instead of the full 20 letter alphabet. Previous studies have shown that a reduced grouping of alphabets based on statistical properties accurately capture the information provided by the full 20 letter alphabet set while increasing the computational efficiency.[12,19,60] For example, a mutation L90M from L(Leucine) → M(Methionine) at position 90 in HIV-1 protease (99 residues long) has a probability $\binom{99}{1} e^{\Delta E^{90}_{\mathrm{L}\to\mathrm{M}}}$ associated with the mutation each KMC step. Figure 1 shows the schematic for the KMC process starting from a single seed sequence.

When run for many steps, these trajectories reach Markov equilibrium satisfying the detailed balance conditions, $P(S_1) f_{S_1\to S_2} = P(S_2) f_{S_2\to S_1}$, consistent with the definition from the Potts model above that $P(S) \propto e^{-E(S)}$. Thus, if an ensemble of parallel trajectories is run for long enough to reach equilibrium, then the final sequence from each trajectory the ensemble will follow the distribution $P(S)$, which also describes the distribution of sequences in our seed sequence data set. Since our seed sequences reflect an equilibrium state, all of our ensemble averages reflect equilibrium averages.

The effect of HIV evolutionary kinetics is illustrated by studying the time-dependent contingency and entrenchment effects using the KMC simulations quantified as the average contingency and entrenchment score $C_{\mathrm{M}}(0) = 0$ which is formulated as

$$\langle C_{\mathrm{M}}(\tau)\rangle = \langle \Delta E_{\mathrm{M}}(t = \tau)\rangle - \langle \Delta E_{\mathrm{M}}(t = 0)\rangle \tag{5}$$

where, $\langle C_{\mathrm{M}}(\tau)\rangle$ is the average contingency and entrenchment score for mutation M, $\langle \Delta E_{\mathrm{M}}(\mathrm{t} = \tau)\rangle$ is the average effect of the mutation (M) in background at time $t = \tau$, and $\langle \Delta E_{\mathrm{M}}(\mathrm{t} = 0)\rangle$ is the average effect of the mutation (M) in background at time $t = 0$ when the mutation occurred.

The detailed process of evaluating the contingency and entrenchment score $\langle C_{\mathrm{M}}(\tau)\rangle$ for a specific drug-resistance mutation (L90M) in PR HIV-1 is discussed in section 3.2.

2.4. Restricted Potts Models. In order to explore the collective effects of epistatic interactions on evolutionary

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

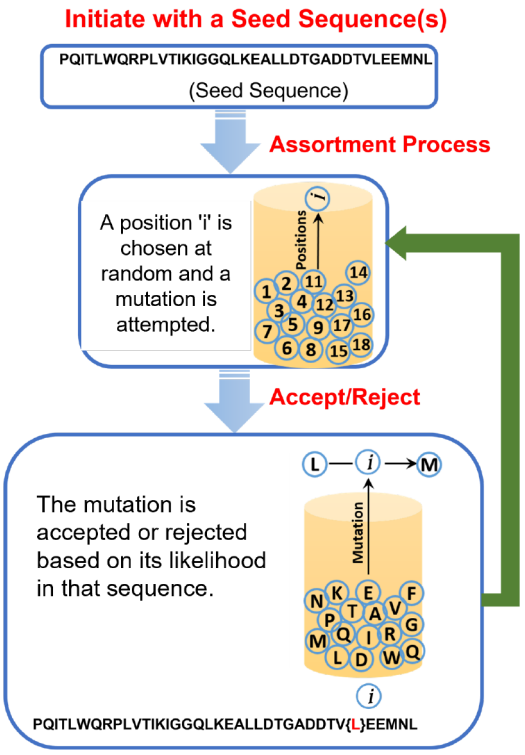


Figure 1. Schematic diagram of the metropolis algorithm which is used during the kinetic Monte Carlo simulation to model the evolution process of HIV sequences.

kinetics, we construct "restricted" Potts models in which some positions are epistatically "decoupled" from the rest, in order to observe the effects of this decoupling on evolutionary dynamics. To construct a restricted Potts model in which a position only covaries with a limited number of other positions, we first determine the "least strongly" coupled positions to a chosen focal position. As a measure of overall coupling strength in our epistatic model between two positions i,j, we use a standard measure, the Frobenius norm, defined as

$$F_{ij} = \sum_{\alpha\beta} (J^{ij}_{\alpha\beta})^2 \tag{6}$$

where the model parameters $J^{ij}_{\alpha\beta}$ are evaluated in the "zero-mean" gauge as described in previous publications.[50,61] We then remove columns with low F_{ij} from the MSA, infer a Potts model on this reduced MSA, and generate an ensemble of trajectories using out KMC scheme.

We then evaluate the autocorrelation of the mutant indicator function for this ensemble of trajectories. The mutant indicator $I_m(t)$ is 1 at times t that the sequence has the mutant m, and 0 otherwise. The autocorrelation of the mutant indicator function is evaluated as

$$\rho_I(\tau) = \frac{\langle (I_m(t) - \overline{I}_m)(I_m(t+\tau) - \overline{I}_m)\rangle}{\langle (I_m(t) - \overline{I}_m)^2\rangle} \tag{7}$$

where $\overline{I}_m = \langle I_m(t)\rangle = P^m(S)$.

3. RESULTS AND DISCUSSION

3.1. Epistasis Causes Amino Acid Biases in the Drug-Experienced State. As a consequence of epistasis, the effect of fitness of a mutation, at the time of its introduction, depends on the presence of specific patterns of other mutations in the genomic sequence background;[62−65] the likelihood of fixation of this mutation then depends on the preceding history of other mutations, a phenomenon known as "contingency" in evolutionary biology.[66−70] Here, we will denote this mutation as the "focal" mutation of interest, and other mutations in the sequence background as "accessory". After the focal mutation occurs, it will increase the likelihood of certain subsequent mutations at other sites which themselves increase the fitness of the focal mutation, and reversion of the focal mutation can become increasingly deleterious with the passage of time, thus, leading to an evolutionary "entrenchment" of the focal mutation.[21,71] Mutations that become entrenched, although typically nearly neutral at the time the mutation occurs, would be highly deleterious in the absence of the preceding mutations, and conversely once entrenched, can become increasingly deleterious to revert over time.

Here, we explore the phenomena of contingency and entrenchment in the context of evolution of drug resistance in HIV. We model the epistatic effects for HIV drug-target proteins using a Potts model, inferred from the observed mutational covariation patterns in large multiple sequence alignments obtained from patient viral samples, where the model parameters represent pairwise epistatic interaction strengths between all residues at all position-pairs (see section 2). The Potts model provides a log-likelihood function $E(S)$ for each sequence S which is interpreted as a proxy for fitness; sequences with higher log-likelihoods have higher fitness and *vice versa*. Thus, $E(S)$ defines an epistatic "fitness landscape", giving a fitness for every possible sequence. Critically, this model is "generative", as one can generate new sequences from the model (by very long Monte Carlo trajectories) which reproduce the mutational covariation statistics of the data set.[72] Our multiple sequence alignment data set is composed of viral sequences from patients under drug treatment, collected in the Stanford HIV Database (see section 2); therefore our inferred Potts model represents the HIV fitness landscape in a drug-exposed viral environment.

The favorability of a mutation occurring at a particular position in a given sequence background S can then be evaluated by the change in the Potts statistical energy[14,19] for acquiring that specific mutation from the wild-type residue (eq 8).

$$\Delta E(S^i_{\alpha\to\beta)}) = E(S^i_\alpha) - E(S^i_\beta) \tag{8}$$

where $\Delta E = 0$ indicates the neutral situation where a focal mutation is neither favored or disfavored by the background, i.e., the sequence with the mutant has equal fitness as with the wild-type residue at that position. We define a small range $(-0.5 < \Delta E < 0.5)$ near the neutral point $(\Delta E = 0)$ 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 favorabilities in different sequences.

We choose the primary drug-resistance mutation (DRM) L90M in HIV protease as an illustrative mutation to depict the epistatic effects of the sequence background. L90M leads to drug resistance toward a variety of ART drugs (SQV, NFV, IDV, and LPV).[73,74]Figure 2 shows the predicted favorability of the mutation L90M as a function of the number of

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

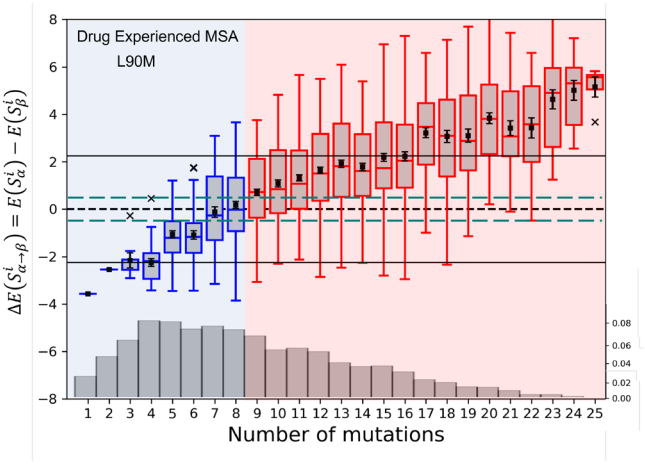


Figure 2. Effect of epistasis and entrenchment on the favorability of a primary resistance mutation. The Potts favorability (ΔE) for a focal mutation L90M (y-axis) is evaluated for all sequences in our data set, and the distribution is shown as box plots annotating the first, second, and third interquartile ranges conditional on the total number of other mutations in each sequence with respect to wild-type consensus (x-axis). Whiskers extend to 1.5 times the interquartile ranges with outlier sequences marked as × symbols, and the mean values are marked as black square boxes. The black dashed line indicates $\Delta E = 0$, and the area around $\Delta E = 0$ is indicated with green dashed lines ($-0.5 < \Delta E < 0.5$) to define the neutral zone. Sequences whose energy difference fall above the neutral zone are defined as entrenching backgrounds favoring the mutation. The mutation L90M becomes favorable on average when there are about more than 9 mutations with respect to wild-type consensus (red box plots). The black solid lines represent $\Delta E = \pm 2.5$ (1 standard deviation around $\Delta E = 0$). The distribution of the total number of mutations with respect to the HIV-1 subtype B wild-type consensus sequence within the drug-experienced patient population in Protease (PR) are shows at the bottom of the box plot, gray. A version of this figure has been shown in previous studies.[14,19]

mutations (Hamming distances)[14,19] relative to the HIV-1 subtype B consensus sequence, for drug-experienced sequences found in the Stanford HIV drug database.

For sequences with fewer mutations away from the wild-type consensus, the L90M mutation is on average highly disfavored. However, for sequences with greater numbers of accessory mutations, the primary mutation L90M is increasingly favored, eventually becoming entrenched when $\Delta E > 0.5$ when there are 9 or more mutations on average. However, it is not just the number of mutations but "which" mutations are present (Figure 2 whiskers in boxplots), that determine the entrenchment of the primary mutation, as even for a fixed number of accessory mutations we observe large variability in the favorability of the primary mutation.

We emphasize that Figure 2 does not explicitly show a time-ordering of sequences, as this figure shows independent sequences sampled from different patients undergoing drug therapy, rather than a set of sequences sampled from a single patient over time.

However, it is commonly observed that a single host's HIV population consensus sequence is initially more close to wild-type HIV before drug exposure,[1,75,76] but after drug exposure over time additional DRMs are accumulated which leads to ART failure.[56,57] In the next section we establish a concrete model for such time-course kinetics.

3.2. Contingency and Entrenchment of Drug-Resistance Mutations in HIV. We now develop a model of evolutionary kinetics over this fitness landscape. We model sequence evolution under drug-selection pressure using the KMC method with a drug-experienced Potts Hamiltonian fitness landscape. In each of a series of time-steps, we attempt mutations at randomly chosen positions and accept mutations based on the change in Potts energies, using the Metropolis criterion, as described in methods. By this scheme we model the evolutionary process as a kinetic Markov process where the Markov states are equivalent to sequences, and the transition rates are functions of Potts ΔE. When run for many steps until Markov equilibrium (see section 2), this method generates new sequences whose mutation frequencies and covariation statistics match those of the observed patient sequences the Potts model was built on.

Using these kinetic simulations, we first illustrate and quantify the effect of epistasis in HIV evolutionary kinetics by demonstrating the time-dependent contingency and entrenchment effects defined in the previous section (section 2.3). This effect becomes apparent by averaging over DRM events in

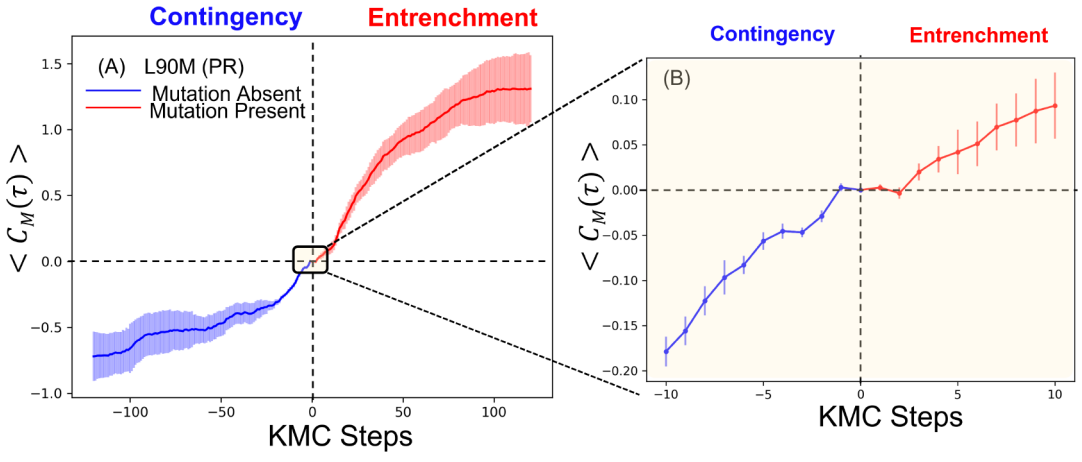


Figure 3. Contingency and Entrenchment of the primary mutation L90M as a function of KMC steps. Plot (A) shows the effect of contingency and entrenchment quantified as the contingency and entrenchment score ($C_M(\tau)$) of the primary drug-resistance mutations L90M as a function of the Kinetic Monte Carlo steps. Plot (B) is the zoomed panel of plot (A) near the neutral region ($\tau = 0$) where we can see more variation of contingency and entrenchment score ($C_M(\tau)$). The variance is plotted with vertical lines at each KMC step in plots A and B.

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

many independent long-running trajectories which have reached Markov equilibrium as defined above. This averaging eliminates effects due to nonequilibrium starting conditions rather than due to epistasis, for instance if the initial sequences have very low favorability for the mutation which can then only rise. In Figure 3, we demonstrate contingency and entrenchment in the KMC HIV trajectories, by evaluating the relative background-dependent bias for a focal DRM as a function of time relative to the moment that DRM occurred, averaged over many DRM mutation events.[12,14,21,69,70] We choose the primary DRM L90M in HIV protease (PR) to depict the epistatic effects of the sequence background as it evolves over time, in accordance with ref 14, showing that epistatic interactions play a stronger role determining the fitness landscape of HIV proteins affected by DRMs compared to non-resistance-associated mutations (Figure 3).

In detail, for each trajectory we find all time windows centered such that a reference time ($t = 0$) is a KMC step in which the focal mutation *L90M* occurred, and the time-window extends forward and backward 120 KMC steps from that reference time. Overall, the average contingency and entrenchment score $\langle C_M(\tau)\rangle$ plotted in Figure 3 is calculated by averaging over all such time windows for each trajectory and over all trajectories. We compute the bias or the favorability for the focal DRM (L90M mutation), ΔE, at each point within each time window and subtract the ΔE at the reference time $t = 0$ from each window giving an entrenchment score $C_M(\tau)$, so that the value at $C_M(0) = 0$ for all windows. In other words, the entrenchment score $C_M(\tau)$ is the difference between the favorability of L90M at time t relative to the favorability of L90M at the moment it occurred at $t = 0$. As the value of $C_M(\tau)$ increases from zero it represents the increasing likelihood or favorability of that DRM at time t relative to the occurrence reference point ($t = 0$). On the other hand, more negative $C_M(\tau)$ values shows the disfavorability of that DRMs at that point of time compared to the reference point ($t = 0$).

Before the mutation event, as other mutations are gradually accumulated, changes in the background reduce the bias for the wildtype residue, allowing the focal L90M substitution to occur (Figure 3, blue color). The average contingency and entrenchment score $\langle C_M(\tau)\rangle$ gradually increases from ca. −1.0 to 0, and the fitness costs for the focal mutation decrease from an initial $\langle \Delta E\rangle = -1.09$ until the mutational event occurs at nearly neutral ($\langle \Delta E\rangle = -0.34$ at $t = 0$) cost to fitness ("contingency"). After the mutational event, (Figure 3A,B, right side, colored with red) a "back reaction" of the sequence background due to the focal L90M mutation increases the likelihood of compensatory mutations at coupled positions, and these mutations accumulate to entrench the focal mutation, with a mean fitness benefit $\langle \Delta E\rangle = 1.25$. The favorability of the focal mutation continues to increase as the mutation becomes "entrenched" in its background with reversion being increasingly disfavored.

A non-DRM A71T in PR is also studied to analyze the contingency and entrenchment effect on non-DRMs. We found that the contingency and entrenchment effect is less pronounced in A71T (Figure S1) compared to L90M. Similarly we have extended our study to DRMs in RT. We have studied the contingency and entrenchment behavior of the K65R drug-resistance mutation in RT. Like PR, RT mutations show similar contingency and entrenchment effects (Figure S2). Overall, the contingency and entrenchment effect is a general feature of evolving mutations irrespective of the particular protein. The characteristics of the contingency and entrenchment effects can differ significantly depending on the environment, e.g., the strongest effect observed for DRMs compared with non-DRMs.

Gupta and co-workers showed that the most entrenched mutations are the ones which are at some local maxima in the fitness landscape and accumulating correlated mutations as observed in Figure 3, can unlock pathways to these local fitness maxima.[77] The local maxima can be 100 times more favorable in particular backgrounds, and these highly entrenching sequences pose a significant risk for the transmission of drug resistance and in the persistence of drug-resistant viruses. Therefore, in the context of patient populations, a study of the dynamics (autocorrelation decay) of entrenched sequences is very relevant, and in this section, we study the equilibrium dynamics of such highly entrenching sequences within the drug-experienced ensemble.

3.3. Fluctuations in the Background-Dependent Evolutionary Constraint Reveal Two Time Scales. Having demonstrated the existence of a kinetic trapping effect for HIV DRMs, we next consider the question: What are the relative time scales over which this trapping effect fluctuates? To explore this question, we study the autocorrelation function of the Potts DRM relative biases (ΔE), which reflects the kinetic behavior of the background coupling and the time required to "forget" the fluctuation of the background energy difference ΔE from its equilibrium average $\langle \Delta E\rangle$ due to the dynamics associated with a focal mutation's interactions with its background.

The relative time scales over which background trapping fluctuates is of interest because we expect that the mechanisms which determine these time scales are related to those that determine clinical time scales for drug escape as well as reversion of DRMs on removal of therapy or after transmission to a new host. It has been shown in previous studies[78,79] that large differences in escape times for the virus to evade host immune pressure targeting the same epitope in different patients can be explained by differences in patient viral background mutations. Likewise, for drug-pressure escape mutations in response to antiretroviral therapy (ART) it has been suggested[14,19] that many of the most strongly entrenched drug-resistance mutations in HIV proteins revert slowly in drug-naive patients with transmitted resistance or in drug-experienced patients after withdrawal of ART.[80−83] These trapping effects can cause DRMs to be 100 times more favorable in some backgrounds than others, and such highly entrenching backgrounds pose a significant risk for the transmission of drug resistance, and in the persistence of drug-resistant viruses.

We evaluate the decay of the autocorrelation function in order to characterize how, on average, the focal mutation evolves on the fitness landscape. As a sequence evolves, the favorability of a particular potential mutation as measured by ΔE will vary due to epistasis, which we annotate as a function of time as $\Delta E(t)$. At long times, this favorability will fluctuate around a mean value $\langle \Delta E\rangle$ corresponding to the log likelihood of the mutation relative to the wildtype residue at equilibrium which is different for different mutations, where $'\langle ...\rangle'$ represents ensemble averages. For a particular mutation, we measure the fluctuations from the average favorability as $\delta\Delta E(t) = \Delta E(t) - \langle \Delta E\rangle$. The autocorrelation of $\delta\Delta E(t)$, i.e., the correlation of the fluctuations at a reference time 0 relative

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

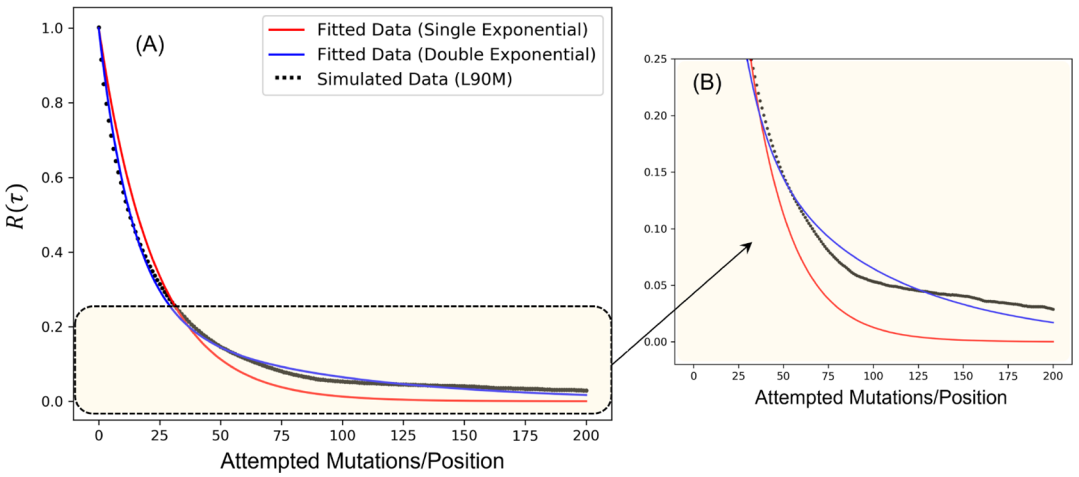


Figure 4. Decay of the function $\delta\Delta E$ autocorrelations in the drug-experienced ensemble with the function of the average number of mutations attempted per site depends on the length of the protein (99 in PR) in HIV-1 Protease. (A) Decay of the fluctuations of the $\delta\Delta E$, $\delta\Delta E$ autocorrelations as a function of the number of mutations attempted at a site on average for primary resistance mutations L90M in HIV-1 PR, are fitted with single exponential (red) and double exponential (blue) fitting, respectively. (B) Zoomed in panel of (A) where it can be seen that the $\delta\Delta E$ autocorrelation function is better fitted with double exponential decay than single exponenetial decay.

Table 1. Autocorrelation Function Decay Times and KMC Steps Associated with Decay Times, Amplitudes, and Coefficients of Determination (R^2) of Major DRMs in PR

					R^2	
mutations	time constant (τ_1)[a]	time constant (τ_2)[a]	amplitude (A_1)	amplitude (A_2)	double exponential	single exponential
D30N	82.64	14.08	0.26	0.74	0.99	0.92
V32I	38.27	3.60	0.40	0.60	0.99	0.92
G48V	50.76	8.26	0.33	0.67	0.99	0.94
I54V	181.81	22.22	0.19	0.81	0.99	0.91
L76V	41.66	4.04	0.28	0.72	0.99	0.85
I84V	90.90	9.90	0.26	0.74	0.99	0.83
L90M	76.92	13.69	0.24	0.76	0.99	0.94

[a]The average number of mutations attempted per position (The length of the protein is 99 for PR).

to those at time τ function, will illustrate the time scales over which the background-dependent favorability or constraints vary. The autocorrelation function, $R(\tau)$, is given by

$$R(\tau) = \frac{\langle \delta\Delta E(t)\, \delta\Delta E(t+\tau)\rangle}{\langle \delta\Delta E(t)^2\rangle} \tag{9}$$

Autocorrelation function plots for our focal mutation of interest are shown in Figure 4.

To extract time scales over which the evolutionary constraint varies, it is common to perform an exponential fit to the autocorrelation function, assuming it decays exponentially. For instance, the autocorrelation function for stochastic processes such as the Ornstein−Uhlenbeck process, which model a stochastic variable that fluctuates around its mean value much like $\Delta E(t)$, has an exponential autocorrelation decay. However, we find that the autocorrelation function for $\delta\Delta E(t)$ in our evolutionary kinetic simulations is better described by a sum of two exponential for primary DRMs in PR (see Table 1). We use a double exponential fit, as $Y = A\ e^{(-x/t_1)} + (1-A)\ e^{(-x/t_2)}$ with two distinctive phases (fast and slow) with different decay times (t_1 and t_2) and amplitudes (A and $(1-A)$). In all cases, we find the p-value of an F-test comparing the double-exponential fit to the single-exponential is $p \ll 1e^{-5}$, meaning the double-exponential fit is significantly better. The results of the fit for different DRMs are shown in (Table 1, Figures 4 and S1). This shows that parameters such as ΔE arising from evolutionary kinetics under Potts-like fitness functions are not a simple process like an Ornstein−Uhlenbeck process, which has also been noted in other studies.[84,85]

In Table 1, we find that $\delta\Delta E$ autocorrelation function for some DRMs decay faster than others, for instance time scales for the autocorrelation function for L90M decays are greater than those for V32I and L76V. We hypothesize that mutations with slower decaying $\delta\Delta E$'s autocorrelation function are indicative of mutations whose favorability depends more extensively on the sequence background, which we investigate further below. In our double-exponential fits, we also observe significant differences in two time scales for each DRM (Figure 4, Table 1), and by comparing the amplitude parameters we observe that, the fast time scale is the dominant one. Similarly, the autocorrelation functions of a few DRMs in RT (Figure S4 and Table S1) are also studied to generalize our observations. We have found alike behavior of autocorrelation functions in the DRMs in RT.

We hypothesize that the slower decay time scale is caused by highly entrenched sequences within the population which remain trapped longer, in contrast to the nonentrenched sequences which tend to relax faster. We will further examine this hypothesis below by examining the waiting times for DRMs to occur, and also by examining the autocorrelation decay time of the mutation indicator function, both defined below.

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

3.4. DRM Waiting Times Support a Bimodal Trapping Effect. The trapping effect introduced above has many consequences on DRM kinetics, which we can further illustrate by examining the waiting times for DRMs to occur in simulated trajectories. We evaluate the waiting times for the focal mutation, which we define as the time spans in a trajectory starting from the moment the focal position mutates from wild-type to the focal mutant residue (start point of the waiting time) and then revert back to wildtype (end point of waiting time). We calculate the waiting-time for all DRM mutation events across our ensemble of trajectories, from which we calculate the statistics described next.

A previous study[78] has shown that differences in sequence background mutations of the viral strains in patients can lead to large differences in immune escape times for HIV patients targeting the same epitope. Viruses like HIV subject to antiretroviral therapy (ART) are expected to give analogous results.[14] If the primary resistance mutation is favorably coupled to the background, then the escape mutation will likely take much longer to revert in the population as mutations in the background must occur to reduce the biases favoring the focal mutation, and it has been suggested[14,19] that many of the most strongly entrenched mutations in HIV proteins revert slowly in drug-naive patients with transmitted resistance or in drug-experienced patients after withdrawal of ART.[47,86−88]

To test our trapping hypothesis, we examine the relationship between waiting time and ΔE by evaluating ΔE at the moment the focal mutation occurs, and computing the mean waiting time for focal mutant events with similar initial ΔE. We find that for L90M in PR (Figure 5), longer waiting times are

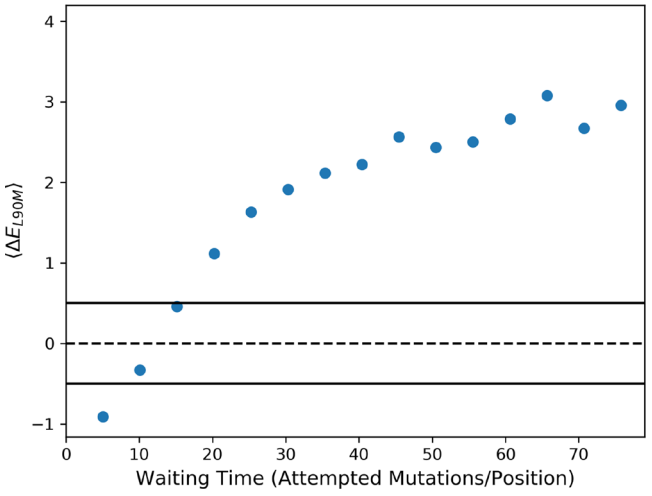


Figure 5. Relation between average ΔE_{L90M} at the moment the focal mutation occurs, with the focal mutant waiting time, in time units of average mutations attempted per site, for the drug-experienced HIV-1 PR fitness landscape. The back dashed line represents the neutral point ($\Delta E = 0$). The solid black lines around $\Delta E = 0$ shows the neutral zone ($-0.5 < \Delta E < 0.5$) of a L90M mutation.

associated with entrenched sequence backgrounds ($\Delta E > 0$). In a previous section (Figure 4), we showed that there are two time scales for the autocorrelation decay of the fluctuations in the Potts energy differences, $\delta\Delta E$s for different DRMs. The shorter and longer time scales associated with L90M are ∼14 and ∼70 attempted mutations per position respectively (Figure 5). Figure 4 shows the relationship between the waiting times to relax back to wild-type residue leucine (L) at position 90 following the mutation to methionine (M) at that position (L90M mutation). For long waiting times (>70 attempted mutation per position, or equivalently 6860 KMC steps), the average background bias $\langle \Delta(E) \rangle$ when the mutation first occurred is ∼ +2.74 corresponding to the sequences which are highly entrenched at the time of mutation. In contrast, for the majority (76%) of waiting times which correspond to short waiting times (<14 attempted mutations per position), the average background bias $\langle \Delta(E) \rangle$ when the mutation first occurred is ∼ −0.62 which favors the wildtype residue leucine at position 90, so it is not necessary for the sequence background to relax in order to promote the transition of the methionine DRM back to the wildtype residue (leucine). This supports our hypothesis (section 3.3) that the slower decay time scale in the fluctuation of $\delta\Delta E$ autocorrelation function is caused by highly entrenched sequences within the population, in contrast to the nonentrenched sequences which decay much faster.

In summary, we have observed that the time autocorrelation function of ΔE as well as the relationship between mutant waiting times and ΔE indicate the presence of two distinct processes at different time scales: the faster reversion of the mutation in nonentrenching sequence backgrounds followed by the slower reversion in entrenching sequence backgrounds. The kinetics of the decay/reversion process is dictated by the nature of mutations and "entrenchment" in the sequence background, with the degree of entrenchment determining how slow or fast the reversion process happens.

3.5. Epistasis Causes Qualitatively Different Waiting-Time Kinetics than Site-Independent Evolution. Epistasis and the trapping effect have other important impacts on kinetics which become apparent by contrast to kinetics under a site-independent model. Analogous to the Potts model, we can infer a nonepistatic "independent" model, given patient sequence data such as used to infer the Potts epistatic fitness landscape, but which differs in that the favorability of a mutation does not depend on the sequence background, and there is no epistasis. Like the Potts fitness landscape, the Independent fitness landscape can be used generatively to produce sequences with the same mutant frequencies as the data set, but unlike the Potts model it cannot capture mutational covariation patterns or epistasis. Here, we fit an Independent model to the same drug-experienced sequence data set as we used to infer the Potts epistatic fitness landscape, and simulate kinetics over both landscapes using our kinetic model, and compare these kinetics. In this section, along with DRMs, we also consider polymorphic mutations in HIV-1 PR. The polymorphic mutations are defined as frequently occurring mutations in viruses that are not exposed to selective drug pressure and they have been known to have very little coupling with other mutations and can be a representation of the independent model.[89,90]

The log probability distribution of waiting times for DRMs and polymorphic mutations in HIV-1 PR within the drug-experienced state, collected across an ensemble of trajectories, is shown in (Figure 6). We observe two qualitative effects: First, in all cases the coupled model on average has longer waiting times than the independent model, showing that epistatic coupling tends to slow the DRM and polymorphic site evolutionary dynamics. Second, the shape of the distribution is different for the coupled versus independent fitness models, with a longer high-waiting-time tail for the coupled model. It is

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

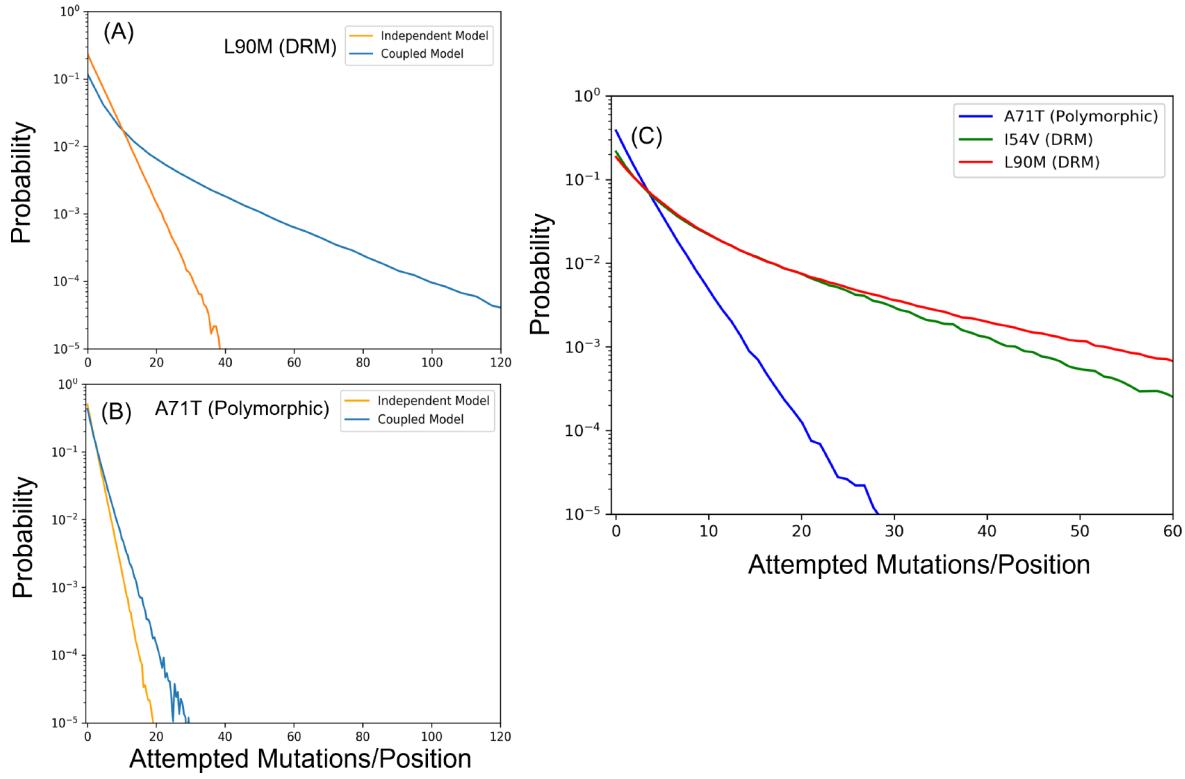


Figure 6. Waiting time (as a function of the average number of mutations attempted per site depends on the length of the protein (99 in PR) in HIV-1 PR) distribution using independent and coupled model. The distribution in plot (A) drug-resistance mutation or DRM (L90M), (B) polymorphic mutation (A71T) in protease (PR). The waiting time distribution with independent model are shown by orange color and with coupled model are shown by blue color. (C) Waiting time distribution of polymorphic mutations (A71T) and two drug-resistance mutation (I54V and L90M) in PR using coupled model.

known analytically that the distribution of waiting times, as we have defined it above, for the independent model follows a single exponential decay, since the favorability (and so mutant acceptance rate) is fixed. A standard way to quantify the deviation of such a process from a single-exponential (Poissonian) waiting process is by the "index of dispersion",[85,91] which is the ratio of the variance $\sigma^2(n)$ to the mean $\mu(n)$ of a related distribution, the number of events n per time-interval, or

$$R = \sigma^2(n)/\mu(n) \tag{10}$$

The distribution of number of events per time interval is computed from the waiting time distribution by choosing a fixed time interval T and given the waiting-time distribution counting how many sequential events would occur in that time span. In other words, we set a clock to "tick" only during moments the focal mutation is present, and we divide time into intervals of size T and measure the mean and variance of the number of mutant events per interval. A key property is that since the independent model gives a single exponential waiting time distribution, its distribution of number of events per time-interval is a Poisson distribution which has $R = 1$, and therefore deviations of the index of dispersion from 1 indicate deviations from site-independent (Poissonian) behavior, due to epistasis.

In Table 2, we show R computed for the coupled and independent models for different focal mutants. We obtain values very close to $R = 1$ in all cases for the independent model as required, with small errors due to computational precision limits. In the Potts coupled model, the deviation of index of dispersion from 1 emphasizes the epistatic coupling between the sites. As the coupling between sites increases, a mutation in a position tends to be more favored by its background and is likely to spend a longer time in the mutant state compared to the independent model which leads to overdispersion (Figure 6A,B). We can interpret the index of dispersion as a quantitative measure of the overdispersion and the strength of coupling of a position with the background.

Table 2. Index of Dispersion of DRMs and Polymorphic Mutation Using Independent and Coupled Model

	index of dispersion	
DRMs	coupled model	independent model
D30N	1.6	1.0
G48V	1.4	1.0
I84V	1.8	1.0
L90M	3.2	1.0
V32I	2.2	1.0
I54V	2.1	1.0
L76V	1.5	1.0
polymorphic mutations	coupled model	independent model
L10T (PR)	1.1	1.0
A71T (PR)	1.1	1.0
V77I (PR)	1.2	1.0

These calculations show that the polymorphic sites have indexes of dispersion close to 1, similarly to an independent model. On the other hand, primary DRM positions have larger R (Table 2). We have plotted the distribution of waiting time of a polymorphic mutation, A71T, and three primary DRMs (I54V and L90M) in Figure 6C. We observe that DRMs like

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

L90M and I54V have higher deviation from an independent model compared to the polymorphic mutation (A71T), as confirmed by their R values. This is also consistent with the decay time observed in Table 1 in the autocorrelation function section (section 3.3) where we found that the favorability autocorrelation functions for L90M and I54V decay more slowly, as expected due to the stronger coupling between the drug resistance position with other compensatory positions, resulting in a mutation becoming favored or entrenched by the background and tending to spend more time in the mutant state. Overall, we find the strength of epistatic effects on the evolutionary dynamics is much larger for DRM positions than polymorphic positions, as expected due to their larger coupling.

3.6. Trapping Effect is Controlled by Many Coupled Positions. The previous results show that epistatic coupling between residues tends to both slow down and qualitatively change the evolutionary dynamics of mutants, and by different amounts for different mutants. To further investigate the mechanisms by which these effects arise, we examine the question of whether this slowing can be best explained through strong coupling of the focal mutation to a small number of other positions or whether it is due to larger scale collective effects of chains and networks of couplings involving many positions. To test this, in this section we gradually decouple the focal position from other positions, and examine the effect on the focal mutant's dynamical time scales.

To decouple the focal position from others we fit a series of Potts models to restricted MSAs in which only certain columns of the original MSA are included. This strategy holds fixed the univariate and bivariate marginals of the focal position, to eliminate time scale biases due to different levels of conservation. In these restricted Potts models, the mutant position can only co-vary with this restricted position set. For each of these reduced Potts model fitness landscapes, we then simulate evolutionary trajectories using our kinetic scheme, and compute the autocorrelation function of the mutant indicator function. As detailed in section 2, the mutant indicator is 1 when the mutant is present and 0 when not present, and the autocorrelation of the mutant indicator function measures the time scales over which the residue at the focal position is "forgotten" and thus gives an overall measure of the time scale of evolutionary dynamics. From each such autocorrelation curve, we use a simple and standard measure of the overall autocorrelation time scale, which is the time at which the autocorrelation becomes less than $1/e$ (0.368), as in the common case of an exponentially decaying autocorrelation curve this will reflect a mean time scale. Starting from the fully coupled model used above in which the focal position can co-vary with all other positions, we gradually remove columns from the original MSA, starting with the positions we determine are the "least strongly coupled" to the mutant position in the full Potts model. We define "least coupled" to be the positions i with lowest Frobenius norm F_{ij} with the mutant position j in the full model (see section 2), which is a standard measure of coupling strength between positions i,j.

We carry out this procedure for the mutants L90M and D30N, which we hypothesized could have different behaviors since D30N is known to co-vary with few other mutations, primarily N88D, while L90M appears to co-vary with a greater number of positions (Figure 7). We recapitulate the overall slowing effect for L90M as its autocorrelation time scale is 2.1 attempted mutations per position (attempts/pos) when it

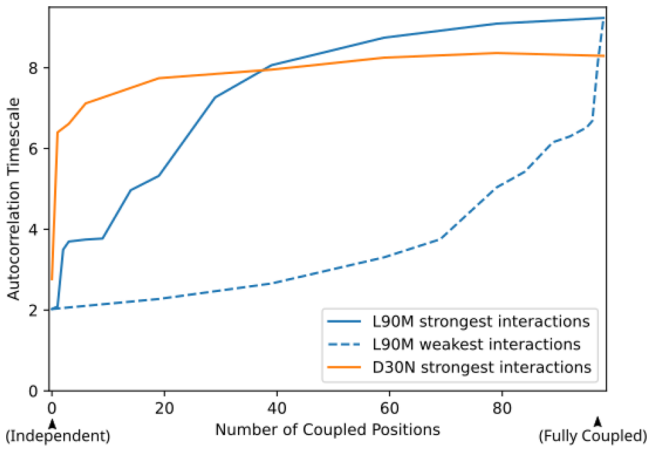


Figure 7. A mutation's evolutionary dynamics speed up as the mutant position is gradually decoupled from other positions. For the mutants L90M (solid blue) and D30N (solid orange), we measure the autocorrelation time scale (see text) of the mutant indicator function, using Potts models built on restricted MSAs in which the mutant position is only coupled to a limited set of other positions. This varies from all positions (the full Potts model, right) to no coupled positions (the mutant position varies independently of all others, left), gradually uncoupling positions from right to left such that the positions in the full model most weakly coupled (as defined in the text) to the mutant position are decoupled first (solid lines). If instead the most strongly coupled positions are decoupled first, then we obtain a different curve, shown for L90M (dashed line).

varies independently from all other positions, which rises to 9.2 attempts/pos when it is fully coupled to all other positions (Figure 7, solid blue line). We also observe that this slowing effect cannot be simply explained as coupling to a small number of other positions: As we couple position 90 to only the top 9 most strongly coupled positions, the autocorrelation time scale rises to 3.7 attempts/pos, much less than the full slowing effect of 9.2 attempts/pos. When coupled to 20 positions, the L90M time scale rises to 5.4 attempts/pos, halfway between the uncoupled and fully coupled time scale, and when 50 other positions are included the time scale rises to 8.5 attempts/pos, or 90% of the increase in time scale to the fully coupled case. Thus, to capture the majority of the slowing effect due to coupling, it appears that collective epistaic effects including a large fraction of the protein must be included, such as 50 positions out of the 99 total positions. While the most strongly coupled positions (top 9) individually contribute more to the coupling effect, together they only contribute a relatively small fraction of the total by these measures. We also test the autocorrelation time scale as position 90 is gradually decoupled from the most strongly coupled positions in the full model, in order to observe the slowing effects of the weakly coupled positions in the absence of the most strongly coupled positions (dashed blue). Similarly to before, we find that while the most strongly coupled positions individually cause the greatest increase in the autocorrelation time scale, as seen in the left end of the plot, they contribute only a minority of the total increase from the uncoupled case. We also observe that the marginal contribution of the top 9 most strongly coupled positions appears larger in this ordering, as the increase due to the 9 in the right end of the dashed line (3.0 attempts/pos difference) is larger than the increase due to the 9 in the left end of the solid line (1.6 attempts/pos difference). This suggests that the slowing effect of the top 9 is increased when

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

covariation with the weakly coupled positions is also allowed, suggestive of a network or chaining effect in which the top coupled positions indirectly couple the focal position to other positions.

In contrast we observe a smaller overall slowing effect for D30N, as the autocorrelation rises from 2.8 attempts/pos when fully uncoupled to 8.2 attempts/pos when fully coupled. When fully uncoupled, the D30N mutant has a higher autocorrelation time scale than L90M because it is rarer, at 7% frequency in the data set as opposed to L90M which has 30% frequency. In the site-independent case in our kinetic model using metropolis kinetics, one can predict analytically that positions with higher conservation will have larger autocorrelation time scales. However, this expected relation between conservation and longer evolutionary time scale which we expect from site-independent variation no longer holds once epistasis is incorporated, and in the fully coupled models the L90M autocorrelation time scale becomes higher than for D30N, demonstrating the importance of modeling epistatic effects. For D30N, we also observe that a smaller number of coupled positions contribute to the slowing effect, as when including the single most coupled position (which is position 88), the time scale of 6.4 attempts/pos is already more than halfway to its maximum of 8.2 attempts/pos, and it reaches 90% of the difference to its maximum value when 20 most strongly coupled mutations are included. This contrast between D30N and L90M shows that there is no universal rule: Some positions like 90 are coupled to larger networks of other positions, while others like 30 are only coupled to a small number.

4. CONCLUSIONS

The evolution of viruses like HIV under drug and immune selection pressures induces correlated mutations due to constraints on structural stability and fitness (ability to assemble, replicate, and propagate infection) of the virus,[92] as a manifestation of the epistatic interactions in the viral genome. It has been shown that long-range epistasis can shift a protein's mutational tolerance during HIV evolution[18] and can make adaptation contingent on evolutionary history.[93] In this study, we follow the dynamics of evolution of drug resistance in HIV through the phases of "contingency" and "entrenchment" as first described in refs 20 and 21 using the changes in Potts energy of a sequence due to a drug-resistance mutation as a proxy for fitness. We show that the entrenchment of primary resistance mutations is in time on the specific pattern of prior changes (mutations) that have accumulated in the sequence background and cannot be simply explained from the number of background mutations alone. This suggests that epistasis plays a major role in the evolutionary kinetics of HIV under drug selection pressure, and both primary and accessory drug-resistance mutations exhibit strong epistatic interactions. Therefore, entrenchment is a likely mechanism by which drug-resistance mutations accumulate and become trapped within the population and in the persistence of drug resistance.

Previous studies of sequence evolution have shown that the longer an amino acid residue has been present at a site, the more deleterious it is to revert and lower is the reversion rate to the ancestral amino acid.[14,19,20,94] This is what is to be expected for DRMs in HIV based on their "entrenchment" in the sequence backgrounds observed in HIV drug-experienced patient sequences[14,19] and in our simulations. In this study, we demonstrated the presence of "contingency" and "entrenchment" effects in "evolutionary equilibrium" over the inferred fitness landscape. To follow the kinetics of drug resistance in the population related to the entrenchment of DRMs, we look at the decay of the time autocorrelation function for the fluctuations in Potts ΔE, which has been shown to be an accurate predictor of the likelihoods of mutations in a given sequence background.[14,19] The decay of the autocorrelation function reveals the presence of two distinct processes at different time scales, that govern the changes in the favorability of a focal mutation. To further investigate the two distinct processses, we then look at waiting times for the Potts ΔE of a mutation in drug-experienced HIV patient protein sequences to reach the ensemble average value, $\langle \Delta E \rangle$, which illustrates that the slower time-scale is dominated by the decay of the mutation from highly entrenching sequence backgrounds, whereas decays from disfavoring but nonentrenching sequence backgrounds dominate the faster decay time-scales. The role of epistatic coupling leading to the trapping effect on the overall kinetics at different sites are analyzed; the epistatic effect is much larger on the drug resistance positions compared with other positions. We also found, by gradually decoupling a DRM from other positions, that the trapping effect in the HIV proteins we tested is often due to covariation with a large number (>20) of accessory positions. This suggests it is often insufficient to consider only one or two accessory positions when evaluating the likelihood of a DRM arising in a patient, and that correlated multiresidue sequence variations play a central role in HIV evolution.

We focused on the equilibrium aspects of DRM kinetics, using ensemble averages over sequence trajectories in a drug-experienced fitness environment. It is ultimately of clinical interest to understand how drug resistance is acquired starting from an initial state consisting of an ensemble of drug-naive sequences, evolving under drug pressure after a drug is administered. Nevertheless, the equilibrium kinetic properties provide important baselines for studying the entrenchment effect, and the equilibrium properties strongly inform non-equilibrium behavior. The entrenchment effect showing the rise in mutant favorability due to an epistatic "back-reaction" with the sequence background is best demonstrated in equilibrium. In a nonequilibrium setting the favorability can rise or fall purely due to a drift toward equilibrium, for instance if the initial sequences were in a highly disfavorable state for the mutant, rather than due to an epistatic phenomenon. Our results using equilibrium averages can then serve as a control in future nonequilibrium studies of epistatic phenomena. In addition, many nonequilibrium properties can be suggested by the equilibrium case. It is well-known in statistical physics, for instance, through "flutuation dissipation" theorems, that the fluctuations in equilibrium of many-body systems such as described by a Potts model, as well as the autocorrelation functions, are related to the response of the system due to external biases (for instance, administration of a drug) and the rate of decay back to equilibrium. We should then expect that the different autocorrelation time scales we measured for different focal mutants will relate to their acquisition rate in clinically relevant nonequilibrium scenarios, and our results again can serve as a reference point in a future study of the nonequilibrium case.

Shah et al.[21] have suggested that even relatively small degrees of epistatic effects (nonadditivity in the stability effects of mutations) can have large effects on the evolutionary process. In this work, we have explored how epistasis affects

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636

the evolutionary kinetics of drug resistance in HIV and found that the role of entrenchment in the sequence background is key to the dynamics of how drug resistance evolves and persists in the HIV patient population. The analysis presented in this study provides a framework to further explore the kinetic pathways through which viral proteins like HIV evolve under drug selection pressure. Our simulation results on epistasis provide a clear direction for the future investigation of the kinetic pathways through which DRMs evolve from being disfavored in the drug-naive HIV patient population to eventually becoming entrenched in drug-experienced patients. Overall, we also find that the Potts model is a powerful classifier that can identify complex sequence patterns that highly favor (entrench) or disfavor each DRM, and their entrenchment is an important factor reinforcing the emergence of drug-resistant viral strains and in the persistence of resistance. Elucidating the dynamics of these epistatic effects for key resistance mutations has the potential to impact the future of HIV therapies, with implications for future drug design strategies to be based on the patient viral reservoir.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c06123.

Contingency and entrenchment plot of a non-DRM mutation (A71T) in PR and a DRM mutation (K65R) in RT; decay of the function $\delta\Delta E$ autocorrelations plots of DRMs in PR (I84V) and RT (K65R); autocorrelation function decay times of important DRMs in RT; frequency of drug-resistance mutations (DRMs) and polymorphic mutations in drug experienced state (PDF)

KMC simulation files (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Allan Haldane − *Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, United States; Department of Physics, Temple University, Philadelphia, Pennsylvania 19122-6008, United States*; Email: allan.haldane@temple.edu

Ronald M. Levy − *Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States; Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, United States*; orcid.org/0000-0001-8696-5177; Email: ronlevy@temple.edu

Authors

Indrani Choudhuri − *Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States; Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, United States*; orcid.org/0000-0002-5829-4625

Avik Biswas − *Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, United States; Department of Physics, Temple University, Philadelphia, Pennsylvania 19122-6008, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c06123

Author Contributions

[‖]I.C. and A.B. contributed equally to this work. A.H. and R.M.L. jointly supervised this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

(1) Coffin, J. M. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **1995**, *267*, 483−489.

(2) 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*, 1582−1586.

(3) 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. *Lancet Microbe* **2022**, *3*, e392−e398.

(4) Cuevas, J. M.; Geller, R.; Garijo, R.; López-Aldeguer, J.; Sanjuán, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol.* **2015**, *13*, e1002251.

(5) Gupta-Wright, A.; Fielding, K.; van Oosterhout, J. J.; Alufandika, M.; Grint, D. J.; Chimbayo, E.; Heaney, J.; Byott, M.; Nastouli, E.; Mwandumba, H. C.; Corbett, E. L.; Gupta, R. K. Virological failure, HIV-1 drug resistance, and early mortality in adults admitted to hospital in Malawi: an observational cohort study. *Lancet HIV* **2020**, *7*, e620−e628.

(6) Pennings, P. S. HIV drug resistance: problems and perspectives. *Infect Dis Rep.* **2013**, *5*, e5.

(7) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286*, 295−299.

(8) Bloom, J. D.; Gong, L. I.; Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **2010**, *328*, 1272−1275.

(9) Haq, O.; Andrec, M.; Morozov, A. V.; Levy, R. M. Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS Comput. Biol.* **2012**, *8*, e1002675.

(10) Deng, N.-J.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the dynamics of HIV-1 protease: a kinetic network model constructed from atomistic simulations. *J. Am. Chem. Soc.* **2011**, *133*, 9387−9394.

(11) Chang, M. W.; Torbett, B. E. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J. Mol. Biol.* **2011**, *410*, 756−760.

(12) 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 Comput. Biol.* **2015**, *11*, e1004249.

(13) Yilmaz, N. K.; Schiffer, C. A.*Antimicrobial Drug Resistance*; Springer International Publishing: Cham, 2017; pp 535−544.

(14) Biswas, A.; Haldane, A.; Arnold, E.; Levy, R. M. Epistasis and entrenchment of drug resistance in HIV-1 subtype B. *eLife* **2019**, *8*, e50524.

(15) Boltz, V. F.; Ambrose, Z.; Kearney, M. F.; Shao, W.; Kewalramani, V. N.; Maldarelli, F.; Mellors, J. W.; Coffin, J. M. Ultrasensitive allele-specific PCR reveals rare preexisting drug-resistant variants and a large replicating virus population in macaques

infected with a simian immunodeficiency virus containing human immunodeficiency virus reverse transcriptase. *J. Virol.* **2012**, *86*, 12525−12530.
(16) Louie, R. H. Y.; Kaczorowski, K. J.; Barton, J. P.; Chakraborty, A. K.; McKay, M. R. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E564−E573.
(17) Shekhar, K.; Ruberman, C. F.; Ferguson, A. L.; Barton, J. P.; Kardar, M.; Chakraborty, A. K. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2013**, *88*, 062705.
(18) Haddox, H. K.; Dingens, A. S.; Hilton, S. K.; Overbaugh, J.; Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **2018**, *7*, e34420.
(19) 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. *Mol. Biol. Evol.* **2017**, *34*, 1291−1306.
(20) Pollock, D. D.; Thiltgen, G.; Goldstein, R. A. Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1352−E1359.
(21) Shah, P.; McCandlish, D. M.; Plotkin, J. B. Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E3226−E3235.
(22) McCandlish, D. M.; Shah, P.; Plotkin, J. B. Epistasis and the Dynamics of Reversion in Molecular Evolution. *Genetics* **2016**, *203*, 1335−1351.
(23) 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*, 678−87.
(24) Biswas, A.; Haldane, A.; Levy, R. M. Limits to detecting epistasis in the fitness landscape of HIV. *PLoS One* **2022**, *17*, e0262314.
(25) Barton, J. P.; Kardar, M.; Chakraborty, A. K. Scaling laws describe memories of host-pathogen riposte in the HIV population. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 1965−1970.
(26) Zhang, T.-H.; Dai, L.; Barton, J. P.; Du, Y.; Tan, Y.; Pang, W.; Chakraborty, A. K.; Lloyd-Smith, J. O.; Sun, R. Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genet.* **2020**, *16*, e1009009.
(27) 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* **2013**, *38*, 606−617.
(28) Matthew, A. N.; et al. Drug design strategies to avoid resistance in direct-acting antivirals and beyond. *Chem. Rev.* **2021**, *121*, 3238−3270.
(29) Henes, M.; Lockbaum, G. J.; Kosovrasti, K.; Leidner, F.; Nachum, G. S.; Nalivaika, E. A.; Lee, S.-K.; Spielvogel, E.; Zhou, S.; Swanstrom, R.; Bolon, D. N. A.; Kurt Yilmaz, N.; Schiffer, C. A. Picomolar to micromolar: Elucidating the role of distal mutations in HIV-1 protease in conferring drug resistance. *ACS Chem. Biol.* **2019**, *14*, 2441−2452.
(30) Rhee, S.-Y.; Taylor, J.; Fessel, W. J.; Kaufman, D.; Towner, W.; Troia, P.; Ruane, P.; Hellinger, J.; Shirvani, V.; Zolopa, A.; Shafer, R. W. HIV-1 protease mutations and protease inhibitor cross-resistance. *Antimicrob. Agents Chemother.* **2010**, *54*, 4253−4261.
(31) Chang, M. W.; Torbett, B. E. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J. Mol. Biol.* **2011**, *410*, 756−760.
(32) Rhee, S.-Y.; Sankaran, K.; Varghese, V.; Winters, M. A.; Hurt, C. B.; Eron, J. J.; Parkin, N.; Holmes, S. P.; Holodniy, M.; Shafer, R. W. HIV-1 protease, reverse transcriptase, and integrase variation. *J. Virol.* **2016**, *90*, 6058−6070.
(33) Levy, R. M.; Haldane, A.; Flynn, W. F. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **2017**, *43*, 55−62. Theory and simulation · Macromolecular assemblies.
(34) Cocco, S.; Feinauer, C.; Figliuzzi, M.; Monasson, R.; Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.* **2018**, *81*, 032601.
(35) Stein, R. R.; Marks, D. S.; Sander, C. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLoS Comput. Biol.* **2015**, *11*, e1004182.
(36) de Juan, D.; Pazos, F.; Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **2013**, *14*, 249−261.
(37) Serohijos, A. W. R.; Shakhnovich, E. I. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* **2014**, *26*, 84−91.
(38) Mézard, M.; Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *J. Physiol. Paris* **2009**, *103*, 107−113.
(39) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67−72.
(40) Ekeberg, M.; Lövkvist, C.; Lan, Y.; Weigt, M.; Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2013**, *87*, 012707.
(41) Sułkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10340−10345.
(42) Ovchinnikov, S.; Kinch, L.; Park, H.; Liao, Y.; Pei, J.; Kim, D. E.; Kamisetty, H.; Grishin, N. V.; Baker, D. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **2015**, *4*, e09248.
(43) Pollock, D. D.; Taylor, W. R. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng. Des. Sel.* **1997**, *10*, 647−657.
(44) McGee, F.; Hauri, S.; Novinger, Q.; Vucetic, S.; Levy, R. M.; Carnevale, V.; Haldane, A. The generative capacity of probabilistic protein sequence models. *Nat. Commun.* **2021**, *12*, 6302.
(45) 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. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E1293−E1301.
(46) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 67−72.
(47) Izopet, J.; Massip, P.; Souyris, C.; Sandres, K.; Puissant, B.; Obadia, M.; Pasquier, C.; Bonnet, E.; Marchou, B.; Puel, J. Shift in HIV resistance genotype after treatment interruption and short-term antiviral effect following a new salvage regimen. *Aids* **2000**, *14*, 2247−2255.
(48) Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *J. Stat. Phys.* **2011**, *144*, 268−302.
(49) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 67−72.
(50) Haldane, A.; Levy, R. M. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Phys. Rev. E* **2019**, *99*, 032405.
(51) Haldane, A.; Levy, R. M. Mi3-GPU: MCMC-based Inverse Ising Inference on GPUs for protein covariation analysis. *Comput. Phys. Commun.* **2021**, *260*, 107312.
(52) Stanford HIV Drug Resistance Database. https://hivdb.stanford.edu/ (Accessed: 09/06/2022).
(53) Foley, B. T.; Korber, B. T. M.; Leitner, T. K.; Apetrei, C.; Hahn, B.; Mizrachi, I.; Mullins, J.; Rambaut, A.; Wolinsky, S.HIV Sequence Compendium; Tecnical Report *Los Alamos National Lab*: Los Alamos, NM, 2018.
(54) Consensus and Ancestral Sequence Alignments Current (Aug. 2004). https://www.hiv.lanl.gov/content/sequence/HIV/CONSENSUS/Consensus.html.

(55) HIV Databases. Los Alamos National Labs, https://www.hiv.lanl.gov (Accessed 09/06/2022).
(56) Rhee, S.-Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298−303.
(57) Shafer, R. Rationale and Uses of a Public HIV Drug-Resistance Database. *Journal of Infectious Diseases* **2006**, *194*, S51−S58.
(58) Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97−109.
(59) Sirur, A.; De Sancho, D.; Best, R. B. Markov state models of protein misfolding. *J. Chem. Phys.* **2016**, *144*, 075101.
(60) Barton, J. P.; De Leonardis, E.; Coucke, A.; Cocco, S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* **2016**, *32*, 3089−3097.
(61) Ekeberg, M.; Lövkvist, C.; Lan, Y.; Weigt, M.; Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *PRE* **2013**, *87*, 012707.
(62) Weinreich, D. M.; Delaney, N. F.; Depristo, M. A.; Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **2006**, *312*, 111−114.
(63) McLaughlin, R. N., Jr; Poelwijk, F. J.; Raman, A.; Gosal, W. S.; Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **2012**, *491*, 138−142.
(64) Natarajan, C.; Inoguchi, N.; Weber, R. E.; Fago, A.; Moriyama, H.; Storz, J. F. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **2013**, *340*, 1324−1327.
(65) Gong, L. I.; Suchard, M. A.; Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2013**, *2*, e00631.
(66) Harms, M. J.; Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **2014**, *512*, 203−207.
(67) Mani, G. S.; Clarke, B. C. Mutational order: a major stochastic process in evolution. *Proc. R. Soc. London B Biol. Sci.* **1990**, *240*, 29−37.
(68) Travisano, M.; Mongold, J. A.; Bennett, A. F.; Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **1995**, *267*, 87−90.
(69) Das, K.; Martinez, S. E.; Arnold, E. Structural insights into HIV reverse transcriptase mutations Q151M and Q151M complex that confer multinucleoside drug resistance. *Antimicrob. Agents Chemother.* **2017**, *61*, e00224-17.
(70) Starr, T. N.; Flynn, J. M.; Mishra, P.; Bolon, D. N. A.; Thornton, J. W. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 4453−4458.
(71) Rhee, S.-Y.; Liu, T.; Ravela, J.; Gonzales, M. J.; Shafer, R. W. Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrob. Agents Chemother.* **2004**, *48*, 3122−6.
(72) McGee, F.; Hauri, S.; Novinger, Q.; Vucetic, S.; Levy, R. M.; Carnevale, V.; Haldane, A. The generative capacity of probabilistic protein sequence models. *Nat. Commun.* **2021**, *12*, 6302.
(73) Craig, C.; Race, E.; Sheldon, J.; Whittaker, L.; Gilbert, S.; Moffatt, A.; Rose, J.; Dissanayeke, S.; Chirn, G. W.; Duncan, I. B.; Cammack, N. HIV protease genotype and viral sensitivity to HIV protease inhibitors following saquinavir therapy. *AIDS (London, England)* **1998**, *12*, 1611−8.
(74) Zolopa, A. R.; Shafer, R. W.; Warford, A.; Montoya, J. G.; Hsu, P.; Katzenstein, D.; Merigan, T. C.; Efron, B. HIV-1 genotypic resistance patterns predict response to saquinavir-ritonavir therapy in patients in whom previous protease inhibitor therapy had failed. *Ann. Int. Med.* **1999**, *131*, 813−21.
(75) Redd, A. D.; et al. Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions. *J. Infect. Dis.* **2012**, *206*, 1433−1442.
(76) Sagar, M.; Laeyendecker, O.; Lee, S.; Gamiel, J.; Wawer, M. J.; Gray, R. H.; Serwadda, D.; Sewankambo, N. K.; Shepherd, J. C.; Toma, J.; Huang, W.; Quinn, T. C. Selection of HIV variants with signature genotypic characteristics during heterosexual transmission. *J. Infect. Dis.* **2009**, *199*, 580−589.
(77) Gupta, A.; Adami, C. Strong Selection Significantly Increases Epistatic Interactions in the Long-Term Evolution of a Protein. *PLoS Genet.* **2016**, *12*, e1005960.
(78) Barton, J. P.; Goonetilleke, N.; Butler, T. C.; Walker, B. D.; McMichael, A. J.; Chakraborty, A. K. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat. Commun.* **2016**, *7*, 11660.
(79) Chen, H.; Kardar, M.Mean-field computational approach to HIV dynamics on a fitness landscape. *bioRxiv (Evolutionary Biology)*, January 11, 2019, 518704, ver. 2. DOI: 10.1101/518704.
(80) Izopet, J.; Massip, P.; Souyris, C.; Sandres, K.; Puissant, B.; Obadia, M.; Pasquier, C.; Bonnet, E.; Marchou, B.; Puel, J. Shift in HIV resistance genotype after treatment interruption and short-term antiviral effect following a new salvage regimen. *AIDS* **2000**, *14*, 2247−2255.
(81) Yang, W.-L.; Kouyos, R. D.; Böni, J.; Yerly, S.; Klimkait, T.; Aubert, V.; Scherrer, A. U.; Shilaih, M.; Hinkley, T.; Petropoulos, C.; Bonhoeffer, S.; Günthard, H. F. Swiss HIV Cohort Study (SHCS), Persistence of transmitted HIV-1 drug resistance mutations associated with fitness costs and viral genetic backgrounds. *PLoS Pathog.* **2015**, *11*, e1004722.
(82) Gandhi, R. T.; Wurcel, A.; Rosenberg, E. S.; Johnston, M. N.; Hellmann, N.; Bates, M.; Hirsch, M. S.; Walker, B. D. Progressive reversion of human immunodeficiency virus type 1 resistance mutations in vivo after transmission of a multiply drug-resistant virus. *Clin. Infect. Dis.* **2003**, *37*, 1693−1698.
(83) Borman, A. M.; Paulous, S.; Clavel, F. Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug. *J. Gen. Virol.* **1996**, *77*, 419−426.
(84) Rizzato, F.; Rodriguez, A.; Laio, A. Non-Markovian effects on protein sequence evolution due to site dependent substitution rates. *BMC Bioinf.* **2016**, *17*, 258.
(85) de la Paz, J. A.; Nartey, C. M.; Yuvaraj, M.; Morcos, F. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 5873−5882.
(86) Yang, W.-L.; Kouyos, R. D.; Böni, J.; Yerly, S.; Klimkait, T.; Aubert, V.; Scherrer, A. U.; Shilaih, M.; Hinkley, T.; Petropoulos, C.; et al. Persistence of transmitted HIV-1 drug resistance mutations associated with fitness costs and viral genetic backgrounds. *PLoS pathogens* **2015**, *11*, e1004722.
(87) Gandhi, R. T.; Wurcel, A.; Rosenberg, E. S.; Johnston, M. N.; Hellmann, N.; Bates, M.; Hirsch, M. S.; Walker, B. D. Progressive reversion of human immunodeficiency virus type 1 resistance mutations in vivo after transmission of a multiply drug-resistant virus. *Clin. Infect. Dis.* **2003**, *37*, 1693−1698.
(88) Borman, A. M.; Paulous, S.; Clavel, F. Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug. *J. Gen. Virol.* **1996**, *77*, 419−426.
(89) Shafer, R. W.; Rhee, S.-Y.; Pillay, D.; Miller, V.; Sandstrom, P.; Schapiro, J. M.; Kuritzkes, D. R.; Bennett, D. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS* **2007**, *21*, 215−223.
(90) Wainberg, M. A.; Brenner, B. G. The Impact of HIV Genetic Polymorphisms and Subtype Differences on the Occurrence of Resistance to Antiretroviral Drugs. *Mol. Biol. Int.* **2012**, 256982.
(91) Cutler, D. J. Understanding the overdispersed molecular clock. *Genetics* **2000**, *154*, 1403−17.
(92) Theys, K.; Libin, P.; Pineda-Peña, A.-C.; Nowé, A.; Vandamme, A.-M.; Abecasis, A. B. The impact of HIV-1 within-host evolution on transmission dynamics. *Curr. Opin. Virol.* **2018**, *28*, 92−101.

(93) Reddy, G.; Desai, M. M. Global epistasis emerges from a generic model of a complex trait. *eLife* **2021**, *10*, e64740.
(94) Naumenko, S. A.; Kondrashov, A. S.; Bazykin, G. A. Fitness conferred by replaced amino acids declines with time. *Biol. Lett.* **2012**, *8*, 825−828.

https://doi.org/10.1021/acs.jpcb.2c06123
J. Phys. Chem. B 2022, 126, 10622−10636