Enhancing Mutational Hotspot Predictions: The Role of Atomic Fluctuations

***Elvis A. F. Martis*[1,3]* *and Evans C. Coutinho*[1,2]**

[1]Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz [E], Mumbai 400098, INDIA.

[2]St John Institute of Pharmacy and Research, Vevoor, Manor Road, Palghar East, District Palghar 401404, INDIA.

[3]**Current Affiliation**: UMR_C 6286, Unité en Sciences Biologiques et Biotechnologies de Nantes, UFR Sciences et Techniques, Nantes Université, Nantes 44322, FRANCE

*Email for correspondence: elvis.martis@univ-nantes.fr

ABSTRACT
Amino acid substitutions in proteins due to gene mutations play a crucial role in the evolutionary process, enhancing an organism's adaptability and survival. The main objective of these substitutions is to improve the protein's functionality and structural integrity. To assess the impact of amino acid changes on protein function, structural stability, and drug binding, various computational and experimental methods are employed. This study aims to explore the relationship between atomic fluctuations and positional variability to identify mutational hotspots. We propose an empirical score that combines atomic fluctuations, measured as "root-mean-squared fluctuations," with amino acid variability, calculated using "Wu-Kabat's coefficient of variability," which can lead to drug resistance. This score has been developed and computationally validated using two well-studied HIV-1 protease inhibitors, nelfinavir and darunavir. Additionally, we applied this method to anaplastic lymphoma kinase (ALK) and its inhibitor, crizotinib, to evaluate its effectiveness. The predictions are accurate for amino acids that directly interact with the inhibitors. One of the advantages of this approach is its computational efficiency compared to free energy-based methods. However, the predictions are less reliable for amino acids that do not directly bind to the inhibitors, and the current methodology does not account for double or triple mutants. Despite these limitations, the simplicity of the method makes it appealing for applications beyond identifying mutational hotspots.

KEYWORDS
Drug resistance, molecular dynamics simulations, atomic fluctuations, amino acid variability, HIV-1 protease, anaplastic lymphoma kinase.

Drug-induced selection pressure[1–3] propels the evolution of infectious organisms and cancer cells towards decreased sensitivity to drugs. TThese organisms or cells adapt through various mechanisms to reduce drug sensitivity. One defensive strategy involves downregulating the expression of influx channels that facilitate drug entry into the cell, leading to lower intracellular drug concentrations. Alternatively, they may upregulate efflux channels that promote the removal of the drug from the cell, further preventing its accumulation. These adaptations are governed by complex pathways that regulate the upregulation and downregulation of proteins responsible for cellular transport. Due to these complexities, this area remains one of the least explored by computational methods that rely on atomistic simulations and free energy calculations.

Target site substitutions[4,5] that disrupt drug binding while maintaining protein function[6,7] represent another mechanism of drug resistance. These substitutions have been thoroughly investigated through computer simulations, which can assess the binding affinity difference between mutant and wild-type proteins. Enzymes within biochemical pathways may undergo amino acid substitutions that enhance the organism's survival. This can occur either by improving protein function via increased catalytic efficiency and stability (accessory mutations) or by evading the inhibitory effects of drug molecules (major mutations). In the latter case, the primary motivation for altering the drug target is to disrupt drug binding while preserving substrate binding. Most computational approaches used to explore the mechanisms of drug resistance focus on understanding the differences in binding patterns between the substrate and the drug molecule, a concept known as the substrate envelope hypothesis[8–10].

Current literature identifies two primary categories of methods for predicting and studying drug-resistant mutations[11,12] The first category consists of sequence-based methods, while the second encompasses structure-based methods. Many structure-based approaches focus on calculating the change in binding free energy of a drug following a mutation\cite{chellappan2007design}. However, methods that compute the absolute free energy of binding are more computationally intensive than those that assess relative free energy changes. The method presented in this study falls within the structure-based category, however, eliminates the need to calculate changes in binding free energy, resulting in reduced time and computational costs. It is widely recognized that drug-resistant mutations tend to arise at sites that are less conserved and exhibit stronger interactions with the drug molecule than with the substrate or natural ligand[8]. This phenomenon helps to minimize or eliminate fitness costs. However, mutations can also occur at sites with stronger interactions with the substrate, where they may enhance catalytic efficiency or overall protein function. Some mutations, on the other hand, can diminish protein function or catalytic efficiency. The likelihood of these mutations being passed on to future generations hinges on the interplay between the fitness costs and the intensity of the selection pressure exerted by the drug. To mitigate fitness costs, compensatory or gain-of-function mutations can restore the protein function.

The core concept of this method involves analyzing binding-site positions through molecular dynamics (MD) simulations conducted on three distinct states: the \textit{apo}-state, the substrate-bound state, and the drug-bound state. Key amino acid residues essential for binding are identified by observing changes in their atomic fluctuations in the presence of the substrate and/or drug. Additionally, the product of these fluctuations and the variability in the amino acid

sequence highlights residues that are likely to tolerate mutations. Multiple sequence alignments further reveal amino acid substitutions that may be acceptable at specific positions. This approach is largely inspired by the work of Kollman et al[13,14], however, it diverges in two significant ways: it substitutes the van der Waals component of binding free energy with changes in atomic fluctuations. Further, it employs a more reliable Hidden Markov Model (HMM) for multiple sequence alignment to enhance the identification of distantly related proteins.

RESULTS AND DISCUSSION

The method described in this work to identify mutational hotspots is fast and simple such that it can routinely be incorporated as a drug-resistance screen in the drug discovery process. The scoring method was first tested on HIV-1 protease inhibitors (viz., nelfinavir and darunavir) whose effects have been measured on clinically observed resistance cases. After the satisfactory performance, the method was used to predict mutations likely to be observed for anaplastic lymphoma kinase by the inhibitor crizotinib. As for structure-based methods that have been applied to study drug resistance, this method can be applied to situations where there is no *a priori* knowledge of drug-resistant mutations. The method can identify residues that are critical for both substrate and drug binding. This information can be harnessed to develop inhibitors that demonstrate strong interactions with residues that lie within the substrate envelope; with lower chances of developing resistance[31–33] in the future.

Equilibration of MD trajectories

It is of paramount importance that atomic fluctuations are extracted from the well-equilibrated part of the MD trajectories. The Cα-RMSD and root-mean-squared average correlation (RAC) (**Figures S1-S4** in supplementary information) were computed to monitor the equilibration of MD trajectories. The maximum in the normalized distribution for Cα-RMSD for all trajectories is seen between ~1.0-2.0 Å. This indicates that the MD-generated conformations do not deviate significantly from their X-ray conformations. Moreover, we also employ RAC, an additional method to evaluate the equilibration of MD trajectories. The rate at which RAC decays suggests the speed at which a system would achieve dynamic equilibrium[34,35]. **Figures S3** and **S4** suggest that RAC tends to decay fast, and after ~50 ns the RAC values are less than ~0.6 Å. Moreover, most of the structural adjustments are seen in the initial parts of the trajectories around ~10-30 ns. Therefore, all per-residue atomic fluctuations were obtained from the last 50 ns of the MD trajectories.

Significance of the ΔFV scores

It is well-established that drug-resistant mutations are more likely to occur at positions that are not well conserved and those that show more pronounced interactions with the drug than with the substrate or the natural ligand[13]. This ensures minimal or no fitness costs. However, mutations are not uncommon at positions that interact more strongly with the substrate, where the mutation is likely to result in an enhanced catalytic efficiency or protein function. Some mutations reduce protein function or catalytic efficiency. The likelihood of these mutations propagating in future generations will depend on the balance between the fitness penalty and the magnitude of the

drug-induced selection pressure. To overcome the fitness cost, compensatory or gain-of-function mutations are known to salvage the protein's lost functions.

We attempt to capture these aspects quantitatively using the ΔFV scores and to locate amino acids where mutations are more likely to be seen causing drug resistance. To understand how drug binding alters atomic fluctuations in the amino acid residues in the wild-type enzyme vis-à-vis the substrate, molecular dynamics simulations for both *apo*- and *holo*-forms were performed in an explicit water box for a sufficiently long time. The drug targets employed here were simulated under three conditions, first, when the binding site is unoccupied (*apo*-protein), second when the substrate occupies the binding site, and third, when the drug inhabits the binding site (*holo*-protein). The residue-wise difference in the atomic fluctuations $\Delta RMSF_{(apo\text{-}substrate)}$ is obtained by subtracting the atomic fluctuations (RMSF) due to substrate binding from those observed in the *apo*-protein (see the first term in **Eq 2a**). This identifies residues that interact with the substrate. Similarly, $\Delta RMSF_{(apo\text{-}drug)}$ detects residues that interact with the drug molecule. Finally, $\Delta RMSF_{(substrate\text{-}drug)}$ identifies residues that are interact more strongly or selectively with either drug or substrate binding (**Eq 2b**); and is computed by subtracting the atomic fluctuations due to drug binding from those observed due to substrate binding. Thus, the ΔRMSF values are very helpful in identifying residues critical in the binding of various ligands. However, it suffers from the limitation that it conveys nothing about how the wild-type residue is likely to be substituted by other residues. To do so, we compute the positional variability in the protein using Wu-Kabat's coefficient of variability (**Eq 1**) from multiple sequence alignment (MSA) of homologous proteins.

The ΔFV scores are the product of ΔRMSF and positional variability (**Eqs 2a** and **b**) and are to be interpreted as follows. A positive value for the ΔFV_1 score indicates that the position is affected by ligand binding, whereas zero or a negative value depicts that the position may not be affected by a ligand. This observation could indicate that there may be no direct interaction between the amino acid residue and the ligand. The magnitude of ΔFV_1 is larger if the position is variable, and smaller if it is conserved. Hence, a large ΔFV_1 indicates that a residue is important for ligand binding. To stress the difference between substrate and drug binding, the ΔFV_2 score has been introduced. If an amino acid position is more important for substrate binding than drug binding, ΔFV_2 is negative, and otherwise it is positive. The magnitude of ΔFV_2 is again depends on the positional variability, for instance, a variable position yields a larger value and a conserved position leads to a small value. Hence, a large positive ΔFV_2 indicates that a residue is important for drug binding and has the potential to be substituted. Therefore, for a position to be a mutational hotspot leading to drug resistance but not interfering with substrate binding, both ΔFV_1 and ΔFV_2 must have positive values. Furthermore, the ΔFV scores can be useful in structure-based drug design to specifically target amino acid residues involved in substrate binding to minimize the chances of drug resistance. For this, positions should be selected that show positive values for ΔFV_1 for the substrate and negative values for ΔFV_2.

Two pitfalls were immediately discernible: first, this method is sensitive for amino acid positions that are in direct contact (within ~5 Å) with the drug/substrate. By contrast, amino acid substitutions that are further away from the binding site will suffer from small ΔRMSF values due to the lack of immediate interaction with the drug/substrate. Substitutions at such positions may still lead to resistance through allosteric effects but are hard to predict with the current

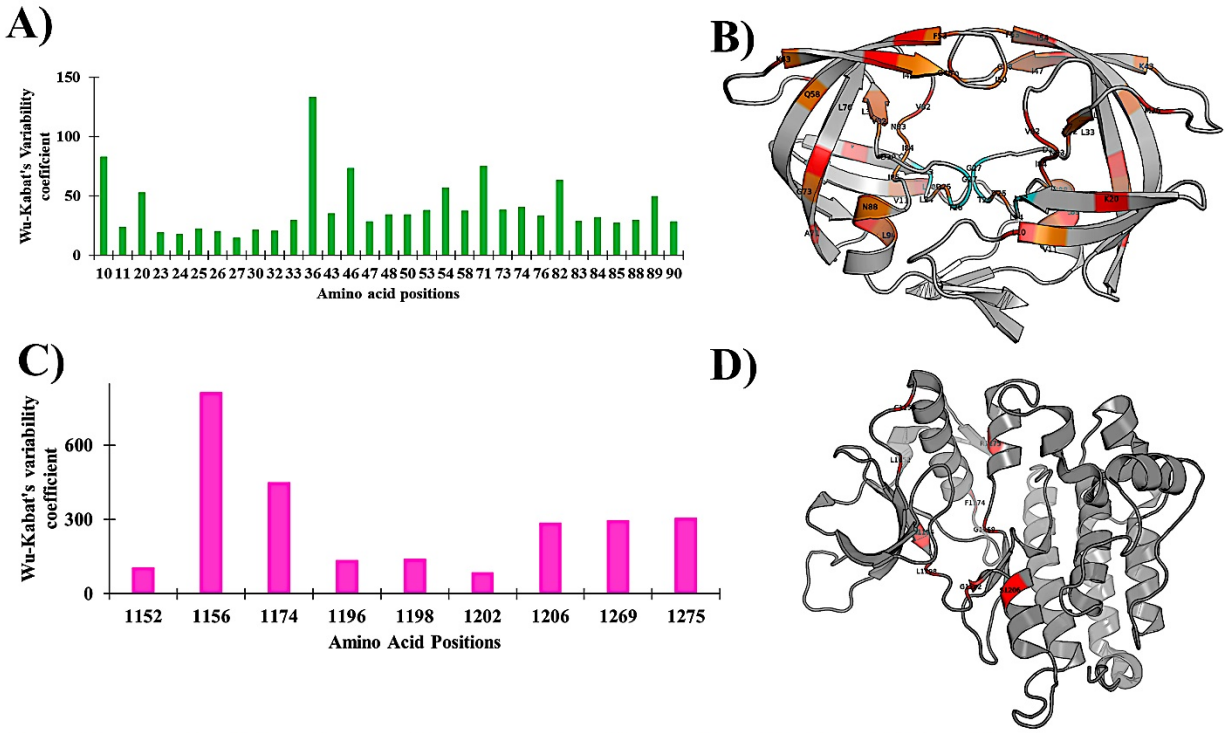
method. Finally, there are no provisions in the present dispensation to address amino acid substitutions at more than one position.

Wu-Kabat's coefficients based on HHBLITS-generated MSA show that most of the positions known to yield resistance are not well conserved.

Before we discuss the outcome of the ΔFV scores, we evaluate the ability of its components, namely the positional variability and the difference in the atomic fluctuations, to enable us to investigate amino acid conservation and identify amino acid positions that interact with drug/ substrate, respectively. **Figure 1** displays the positional variability for HIV-1 protease (single substitutions) leading to drug resistance to any protease inhibitor reported in the HIVdb[36,37] and also for the anaplastic lymphoma kinase inhibitor- crizotinib.

We classify each position in one of three categories based on their amino acid variability – conserved (Wu-Kabat's coefficient of variability < 20), not conserved (> 20), and highly variable (> 40) (**Table S2** Supplementary information and **Figures 1A** and **1B**). Further, we examine the computed positional variabilities for HIV-1 protease, including the residues involved in catalysis - D25(D25')-T26(T26')-G27(G27') (primed residues belonging to chain B of HIV-1 protease). Since D25-T26-G27 are catalytically essential residues, their substitution with any other amino acid will most likely result in an enzyme with compromised catalytic efficiency. This is exemplified by a high frequency (~81%) of these wild-type residues, therefore categorised as conserved positions. However, Wu-Kabat's coefficient of variability for D25 is 22.23, for T26 19.66 and for G27 14.57. Therefore, based on our criteria, D25 must be classified as not conserved, whereas T26 and G27 are labelled as conserved positions. Wu-Kabat's coefficient of variability (**Eq 1**) is not only sensitive to the frequency of wild-type residues but also to the

number of different types of amino acids that occur in the multiple sequence alignment at a particular position in the sequence. In the case of D25, 17 different amino acids are found to substitute the wild-type residue, while it is 15 substitutions for T26 and 11 for G27. Further, we also observe that L23 and L24, where accessory substitutions cause resistance to protease inhibitors, were identified as conserved positions (Supplementary **Table S3A**). This is attributed to the high frequency (~80%) of occurrence of the wild-type residue. Overall, the Wu-Kabat's coefficient of variability computed using HHBLITS-driven multiple sequence alignment (MSA) classifies 26 out of 28 positions (**Tables S3A** and **S3B** and **Supplementary Note 1**) known to cause resistance as not conserved, with two false negative cases (L23 and L24) and one false positive case (D25). Furthermore, in the case of anaplastic lymphoma kinase, all amino acid positions where substitutions occur are highly variable (**Figures 1C and 1D**). Wu-Kabat's coefficients based on HHBLITS-generated MSA show that most of the positions known to yield resistance are not well conserved, which is consistent with the results reported by Kollman *et al*[13].



Cyan- Conserved positions; Orange- Not conserved positions; Red- Highly variable positions

Figure 1. Amino acid variability for HIV-1 protease (for all positions reported in the HIV DB leading to drug resistance to any protease inhibitor reported) and anaplastic lymphoma kinase. **A)** Positional variability computed using HHBLITS (green) for HIV-1 protease. **B)** Depiction of substitution positions on the structure of the protease along with their variability computed using Wu-Kabat's coefficient of variability on MSA generated by HHBLITS. **C)** Positional variability computed using HHBLITS (pink) for anaplastic lymphoma kinase. **D)** Depiction of substitution positions on the structure of the anaplastic lymphoma kinase along with their variability computed using Wu-Kabat's coefficient of variability on MSA generated by HHBLITS. In **B**) and **D**) The cyan-coloured residues have Wu-Kabat's variability < 20 (conserved position), orange-coloured residues have Wu-Kabat's variability > 20 (not conserved position), and red-coloured residues have Wu-Kabat's variability > 40 (highly variable position).

ΔRMSF values are indicators of protein-ligand interactions

We analyze the X-ray crystallographic data to examine whether all amino acid positions mentioned in the HIVdb lie in the active site or outside the active site (**Table S3A** and **S3B**). For nelfinavir-resistant mutations (**Table S4B**), 11 out of 18 positions (in the HIVdb) lie in the active site region, 1 of which is the site for accessory mutations [L23(L23')], 2 positions [D25(D25')

and G27(G27')] are catalytic residues, 5 positions are known to mutate leading to major resistance cases [D30(D30'), I47(I47'), G48(G48'), V82(V82') and I84(I84')]. In the case of darunavir-resistant mutations (**Table S4B**), 12 out of 31 positions lie in the active site region, 1 of which is the site for accessory mutations [L23(L23')], 3 positions [D25(D25'), T26(T26') and G27(G27')] are catalytic residues, 6 positions [V32(V32'), I47(I47'), I50(I50'), L76(L76'), V82(V82') and I84(I84')] are known to mutate leading to major resistance cases. It is worthwhile to note that most of the sites for accessory mutations lie outside the active site/catalytic region, except for L23. Therefore, resistance caused by these mutations is not due to direct interaction with the drug or substrate, and these mutations will not be considered further.

Further, we probe if the parameter, ΔRMSF, can be used to identify residues that show strong/ weak interaction with the drug/inhibitor in question. We only consider positions in the active site of HIV-1 protease within a 5 Å radius from the ligand nelfinavir or darunavir. If an amino acid lies outside this region, the ΔRMSF values are not reliable to ascertain whether the changes in the atomic fluctuations are due to the strength of ligand binding. We then test if there is a relation between the change in mobility of the amino acid and the type of substitution observed i.e., accessory substitution or major substitution, that causes resistance to protease inhibitors. Accessory substitutions do not directly influence the binding of drugs, at least for those considered in this study, however, they are known to improve the stability of the enzyme or compensate for lethal amino acid substitutions[36].

For nelfinavir, positions M46, I54, N88, and L90, which carry substitutions that give rise to drug resistance, lie outside the active site region (Supplementary **Tables S3A** and **S3B**). Substitutions at M46(M46') have led to resistance to the protease inhibitors because of an increase in the

catalytic efficiency, but do not affect inhibitor binding[38–41]. Furthermore, substitutions at L90(L90') are known to reduce the stability of HIV-1 protease[42], which is often compensated for by other substitutions for example, substitutions at N88[43]. Since these mutations lie outside the active site region and do not perturb nelfinavir binding directly, it is beyond the scope of the ΔFV score to show reliable predictions, and therefore, these substitutions will not be looked at further. We note that the probability of side-chain atoms to be found within the active site region is higher than that of backbone atoms (Supplementary **Tables S4A** and **S4B**). Therefore, we use ΔRMSF values of side-chain atoms to compute the ΔFV score.

We evaluate the ability of $\Delta RMSF_{(apo–drug)}$ and $\Delta RMSF_{(substrate-drug)}$ (**Figures 2A** and **2C**) to identify residues that are more critical for drug binding than for substrate binding, which makes them amenable to substitutions that perturb the binding of inhibitors, thus leading to resistance. $\Delta RMSF_{(apo–drug)}$ values indicate that D25(D25'), T26(T26'), G27(G27'), D30', I47(I47'), G48(G48'), V82', and I84' lose mobility upon binding of nelfinavir. It can, therefore, be summarised that $\Delta RMSF_{(apo–drug)}$ can identify all 8 amino acids that interact with nelfinavir which lie in the active site and are thus susceptible to mutation. Positive $\Delta RMSF_{(substrate–drug)}$ values suggest that these 8 amino acid residues (including the catalytic residues) become less mobile in the presence of nelfinavir compared to the substrate, suggesting they show more pronounced interactions with the drug. Nelfinavir shows more prominent interactions with residues in chain B, with exceptions of I47 and G48 where the interactions occur with these residues in both chains A and B. Nelfinavir makes strong interactions with the catalytic residues (D25, T26, and G27), where substitutions are lethal for the virus itself. This is a consequence of the structural

design strategies of protease inhibitors directed in such a way that strong interactions with the catalytic residues occur for efficient inhibition.

Additionally, we investigate 7 amino acid residues (**Table S5A** in Supplementary information) within the active site of nelfinavir, where no substitutions have been reported so far, using $\Delta RMSF_{(apo–drug)}$ values. The $\Delta RMSF_{(apo–drug)}$ values suggest that all 7 amino acids interact with nelfinavir, and more so with these residues in chain B than chain A. However, the $\Delta RMSF_{(substrate–drug)}$ values suggest that 5 (D29, T31, G49, T80 and P81) of these amino acids show more prominent interactions with the substrate than with nelfinavir. Therefore, they should be less prone to substitutions leading to resistance to nelfinavir.

In the case of darunavir, the $\Delta RMSF_{(apo–drug)}$ values suggest that 7 (Supplementary **Table S3B**) amino acids are known to cause major resistance (**Figures 2B** and **2D**). Darunavir interacts with I47, I50, L76, and V82 in both chains as seen in X-ray crystallography[44]. Furthermore, darunavir interacts only with V32' and I84 in chains B and A respectively. The $\Delta RMSF_{(apo–drug)}$ values are in line with X-ray crystallographic observations. We also evaluate the capability of $\Delta RMSF_{(substrate-drug)}$ to identify amino acid positions that show more prominent interactions with darunavir than the substrate. We analyse the 6 amino acids (**Figure 3B**) within the active site known to cause resistance to darunavir. The $\Delta RMSF_{(substrate-drug)}$ values suggest that all positions analysed interact more prominently with darunavir than with the substrate. Thus, we establish that $\Delta RMSF_{(substrate-drug)}$ is efficient enough to identify the residues that interact more prominently with the drug and substitutions at these sites are likely to cause drug resistance.

Further, we investigate the performance of $\Delta RMSF_{(apo–drug)}$ and $\Delta RMSF_{(substrate-drug)}$ to identify the propensity for mutation for 10 positions (**Figures 3A** and **3B**) within the active site region of

Darunavir. Of the 10, three are catalytic residues and for the remaining 7 positions, no substitutions leading to resistance have been reported so far (**Table S5B** Supplementary information). The $\Delta RMSF_{(apo–drug)}$ values suggest that 8 (R8, D25, T26, G27, D29, G49, T80 and P81) out of the 10 residues interact with darunavir. $\Delta RMSF_{(substrate\text{-}drug)}$ values suggest that 7 (R8, G27, D29, T31, G49, T80 and P81) out of the 10 amino acids show more prominent interactions with darunavir than the substrate Therefore, R8, G27, D29, T31, G49, T80 and P81 are amino acid positions more likely to show drug-resistant mutations against darunavir.

It is now established, at least for the systems considered in this study, that $\Delta RMSF_{(apo–drug)}$ and $\Delta RMSF_{(substrate\text{-}drug)}$ are good metrics to identify residues that interact with drugs and are good indicators of the tendency of amino acids to undergo mutation. Therefore, they have been employed as components of ΔFV_1 and ΔFV_2 scores.

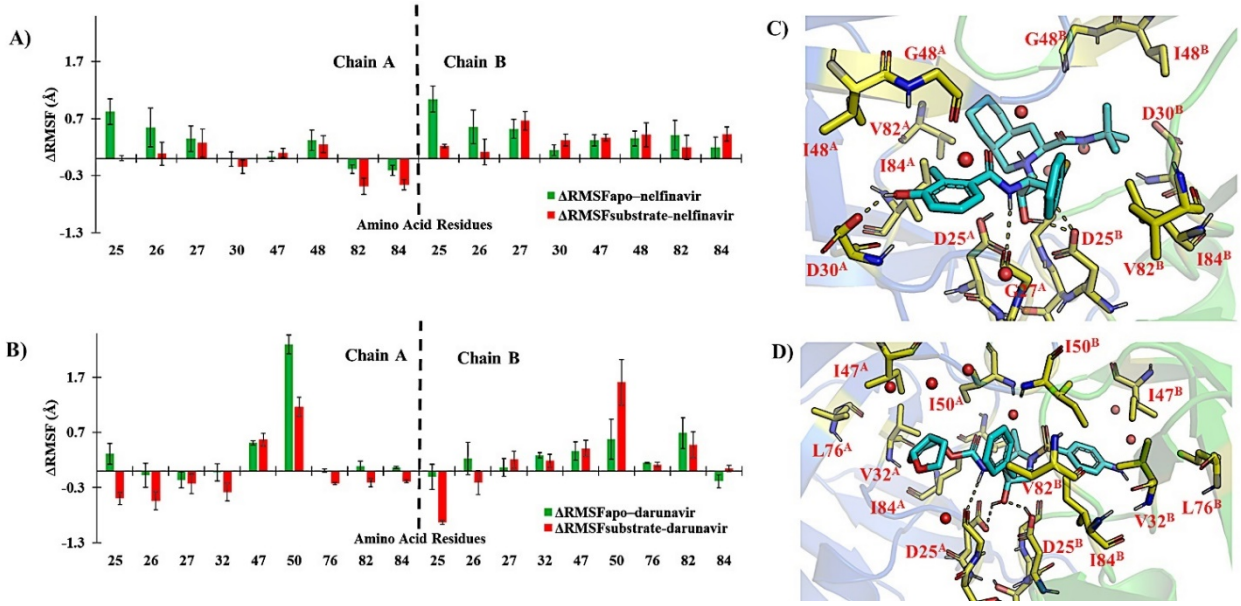


Figure 2. ΔRMSF values (mean ± SEM, n = 3) computed for HIV-1 protease side-chain atoms in the presence of **A)** nelfinavir and **B)** darunavir. **C)** Direct interactions observed in X-ray crystallography of nelfinavir (cyan colour) with HIV-1 protease residues. **D)** Direct interactions observed in X-ray crystallography of darunavir (cyan colour) with HIV-1 protease residues.

Blue-coloured cartoons indicate chain A, green-coloured cartoons indicate chain B, red spheres depict water molecules, and yellow dotted lines indicate hydrogen bonds.

ΔFV scores for HIV-1 inhibitors

The ΔFV score is the product of positional variability and ΔRMSF. While the sign of the score is influenced by ΔRMSF, the magnitude is dictated by both positional variability and ΔRMSF. This is to say, larger values of ΔFV scores most often signify that the position is not well conserved, and therefore, vulnerable to substitution by other amino acids. Furthermore, HIV-1 protease is a symmetric enzyme, however, the inhibitors considered in this work bind asymmetrically, which means the inhibitors show a greater number of interactions than one of the chains of this homodimer. Therefore, ΔFV has been computed for positions on both chains separately; however, for our predictions, only a higher value amongst the two has been considered (**Figure 3**). To investigate the performance of ΔFV scores, we group the amino acid positions into three categories **(Figure 3**), (i) positions within the active-site region known to mutate leading to major drug resistance, (ii) catalytic residues where mutations are known to directly affect the enzyme efficiency and (iii) positions in the active site for which no resistance cases having been reported for the drugs considered in this study. In the case of nelfinavir, ΔFV_1 (**Figure 3A**) shows that all 5 amino acids (D30, I47, G48, V82, I84) that are known to cause resistance, show a positive score that signifies these positions interact with the drug. The ΔFV_1 scores computed for the catalytic positions (**Figure 3A**) show that all 5 positions (D30, I47, G48, V82, I84)) exhibit direct contact with nelfinavir and this is consistent with the X-ray crystallographic data. For positions where no resistance has been reported, ΔFV_1 scores show that all these amino acids physically interact with the protease, and therefore are prone to mutate with nelfinavir, which is

also consistent with X-ray crystallographic data[45]. The ΔFV_2 scores show that all 5 amino acids (D30, I47, G48, V82, I84) are mutational hotspots. Furthermore, the ΔFV_2 scores suggest that even the catalytic residues show a positive value, however, these positions are well conserved, and therefore, not susceptible to mutations. Therefore, the predictions for these catalytic residues are considered false positives. Moreover, for positions (R8, A28, D29, T31, G49, T80, P81) there are no reports as of the date that indicates any resistance to nelfinavir (except T80 and P81), showing positive values for ΔFV_2 which suggests that these positions are prone to mutations. However, the likelihood of R8, A28 and D29 to mutate is low as these positions are fairly conserved with Wu-Kabat's variability coefficient of 19.0, 17.1 and 18.76 respectively. On the other hand, G49 is seemingly a good candidate for mutation as Wu-Kabat's variability coefficient (26.17) suggests that this position is not conserved. A large positive value for ΔFV_2 (27.74) shows that mutation at this position is likely to cause resistance to nelfinavir.

In the case of darunavir, the ΔFV_1 scores show that all amino acids (see **Figure 3B**) directly interact with darunavir. For the catalytic residues, ΔFV_1 shows that all catalytic residues interact directly with darunavir, however, mutations to these residues will likely be lethal to the virus itself. For positions where no resistance cases have been reported, R8, D29, G49, T80 and P81 are seen to interact with darunavir. The ΔFV_2 scores for positions reported to show major resistance to darunavir were predicted to be hotspots for mutations. For the catalytic residues, two out of three amino acids have been identified as non-mutable or mutation with low probability and G27 is falsely classified as positive. Furthermore, for the amino acids where no resistance has been reported, the ΔFV_2 scores indicate that R8, G49, T80 and P81 are more likely to mutate leading to resistance to darunavir.

Therefore, we can conclude that the ΔFV_2 score offers a good indication of amino acids within the active-site region that are prone to mutation, leading to drug resistance.

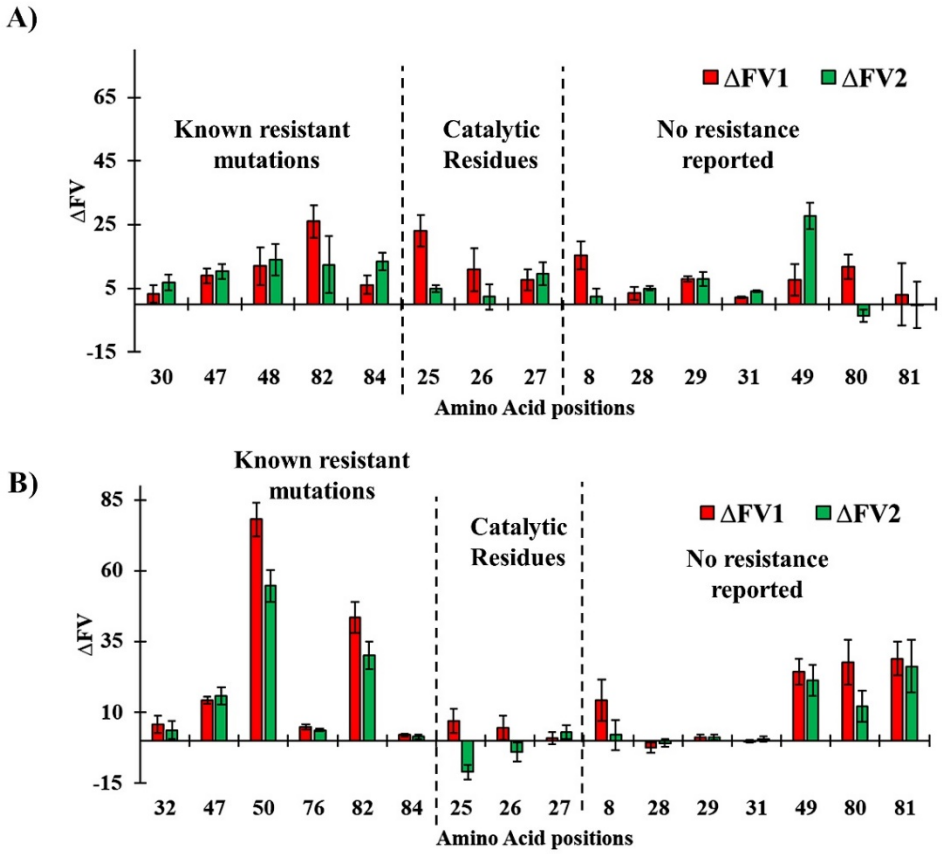


Figure 3: Graphical representation of ΔFV scores for **A)** nelfinavir and **B)** darunavir

ΔFV scores for anaplastic lymphoma kinase inhibitor

As was previously established for the HIV-1 protease system, the differences in the side-chain atomic fluctuations (**Figure 4A**) are more pronounced than those of the backbone fluctuations, and therefore, we use for our analysis atomic fluctuations for side-chain atoms.

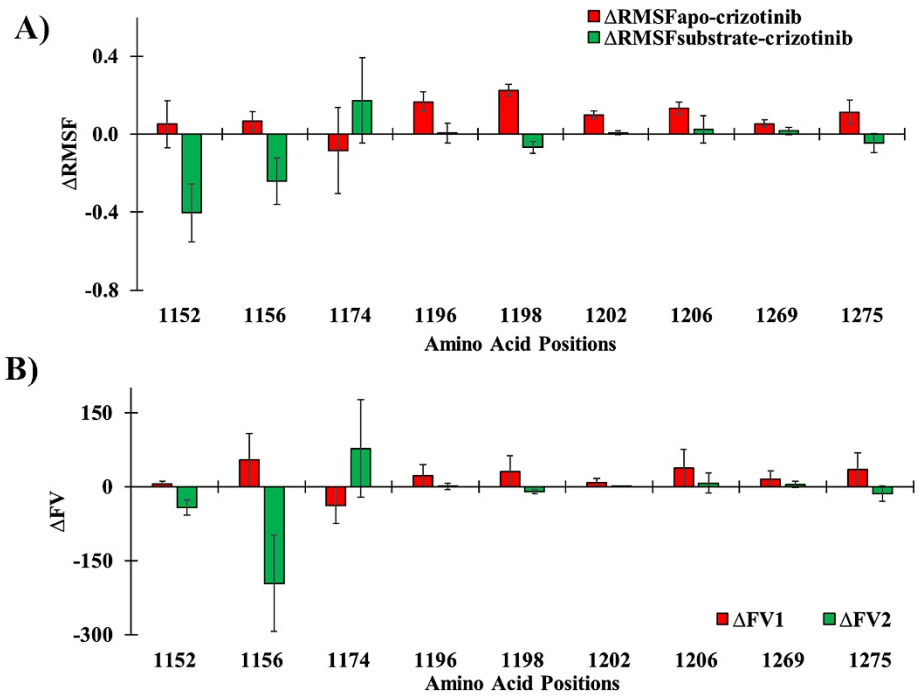


Figure 4. A) ΔRMSF values computed using the side-chain atoms of the ALK kinase domain and **B)** Graphical representation of ΔFV scores for crizotinib.

Since crizotinib is a competitive inhibitor which occupies the same binding site as ATP, the ΔRMSF suggests that ATP and crizotinib have a similar binding pattern with differences observed at L1152 and C1156. The current method shows (**Figure 4B**) only ~38% prediction accuracy in identifying mutational hotspots for crizotinib. This is attributed to the fact that most of the mutations leading to resistance to crizotinib lie outside the active-site region (L1152, C1156, R1275 and F1174) and those that lie within the active-site region (e.g., L1196) are known to enhance the catalytic efficiency of ALK without affecting crizotinib binding. Also, mutations at C1156, F1174, and R1275 are known to increase the ALK function instead of affecting crizotinib binding. This method shows that only G1202, S1206 and G1269 are likely to cause perturbations in the crizotinib binding which is consistent with previous reports[46,47]. Moreover, it has been reported that the L1198F mutant is sensitive to crizotinib[48] and this method shows a

negative prediction for this position based on the side chain data. The current method is nonetheless able to identify those residues (G1202, S1206 and G1269) which affect crizotinib binding and the residues that do not influence its binding are predicted negatively, which is to be expected.

METHODS

Preparation of systems for molecular dynamics simulation

The coordinates for the protein-ligand complexes were retrieved from the Protein Data Bank (PDB); the PDB entries used in this study are listed in **Table S1** in the supplementary information. The X-ray structures were imported and processed in the Maestro-GUI (Schrödinger suite, 2018). All systems were prepared for MD simulations with the *Leap* module in AMBERtools15. Protein atoms were atom-typed according to the AMBER14 (ff14SB) force field.

Generating AMBER-compatible parameters for ligands

All ligands, except ATP, were optimized in vacuum using Gaussian03[15] at the density functional level of theory (DFT) with the B3LYP hybrid functional and the 6-311G^{++}(d, p) basis set. These geometries were used to compute the electrostatic potentials (ESPs) by the Merz-Kollman scheme[16]. To fit the partial charges to the ESPs, the Restrained Electrostatic Potential (RESP) method implemented in *Antechamber*[17] and *parmchk* in AMBERtools15 was used. The General AMBER force field[18] (GAFF) was used to define the atom types, atomic masses, and force constants for bonds, angles, torsions, and improper torsions. All force field parameters for ATP were adopted as defined by Carlson *et al*[19].

Preparing the initial configuration of HIV-1 protease

All mutations induced for stability for crystallographic purposes were reverted to the wild-type amino acid. The substrate for HIV-1 protease considered in this work is the CA-p2 fragment (sequence - KARVLAEAMS). The catalytically inactive D25N mutation was also reverted to the wild-type amino acid. The protonation states of the catalytically important D25 and D25′ residues were adopted as reported by Wright *et al*[20].

Preparing the initial configuration of human anaplastic lymphoma kinase (ALK)

In the X-ray crystal structures of human anaplastic lymphoma kinase (PDB id 2XP2 and 3LCT), coordinates have not been reported for amino acid residues from L1122 to E1129. To correct this, Modeller's loop refinement tool in the Chimera-GUI[21] was used to predict the missing loop. ATP was built by adding a phosphate group to the coordinates of ADP (PDB id 3LCT). The initial configuration for the *apo*protein was built by deleting the ADP molecule from 3LCT (PDB id).

Molecular dynamics simulations

All complexes were solvated with TIP3P waters[22] enclosed in a box with dimensions extending 12 Å beyond any protein atom, and an appropriate number of neutralizing ions were added to maintain the neutrality of the system. The complexes were relaxed using three cycles of minimization of 50,000 steps each. The first 10,000 steps of minimization adopted the steepest descent algorithm, and the remaining 40,000 steps used the conjugate gradient algorithm. The cut-off for evaluating non-bonded interactions was placed at 8.0 Å for the minimization phase.

The first cycle of minimization was performed with 25 kcal/(mol Å^2) restraining force on the solute atoms, this force constant was reduced to 5 kcal/(mol Å^2) in the next cycle, and the final cycle was performed without any restraining forces on either the solute or the solvent atoms. The system was heated from 0 K to 300 K for 1 ns with 5 kcal/(mol Å^2) restraining force on the solute atoms using the Langevin dynamics, with a collisional frequency of 2 ps^{-1}. After heating, the system was equilibrated by MD simulations under the NVT ensemble for six cycles of 1 ns each, where the restraining force on the solute atoms was gradually reduced from 5.0 to 0.0 kcal/(mol Å^2). The next step employed the NPT ensemble for density equilibration for another six cycles of 1 ns each, with the restraining force on the solute atoms gradually decreasing from 5 to 0 kcal/(mol Å^2). In the equilibration phase, the cut-off for evaluating non-bonded interactions was placed at 9 Å. In the production phase of the MD simulations, all systems were simulated for 100 ns in triplicate, amounting to a 300 ns trajectory for each system. The long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) summation method[23]. The minimization and equilibration phases of MD simulations were carried out with the MPI version of the *SANDER* and *PMEMD* codes implemented in AMBER14[24], respectively. The production phase was carried out using the CUDA-enabled *{Le Grand, 2013 #4}*code of *PMEMD* in AMBER14. All simulations were performed with the SHAKE[25] algorithm to constrain bonds to hydrogen atoms, with the tolerance set to 0.00001 Å. A 2 fs integration time step was used to solve Newton's equations of motion. For analysis, the conformations were saved every 100 ps, amounting to 1000 snapshots per trajectory (3000 per system).

Analysis of MD trajectory

Analysis of the MD trajectory was performed with the *Cpptraj*[26] module in AMBERTools15. Before the analysis, all solvent atoms and neutralizing ions were stripped off, and the protein conformations were aligned to the first frame. The mass-weighted Cα root-mean-square deviation (RMSD) for the protein was computed using the X-ray conformation as the reference. The convergence of MD trajectories was estimated using the root-mean-squared average correlation (RAC) with mass-weighted RMS averaging of Cα atoms, with the first frame of the trajectory as the reference. The root-mean-square fluctuations (RMSF) per residue for the backbone and side-chain atoms were computed from the last 50 ns part of the trajectories.

Computing amino acid variability and ΔFV score.

The gapped multiple sequence alignment (MSA) to compute the amino acid variability was generated using the program *HHBLITS*[27]. *HHBLITS* uses Hidden Markov Models (HHM) to search for remote homologous proteins. The E-value was set to 0.0001, a maximum of 8 iterations was performed, and the UNICLUST30[28] database (default) was used to search for homologous proteins. The MSAs were imported into *Bioedit* (biological sequence editor) to compute the amino acid variability at each position. In this work, the Wu-Kabat variability coefficient[29] (**Eq 1**) was computed at each position to measure the positional variability.

$$(\text{Wu} - \text{Kabat variability coefficient})^i = \frac{(N * k^i)}{n^i} \qquad \textbf{Eq 1}$$

where N is the total number of sequences aligned, k^i is the number of different types of amino acids found to substitute the wild-type amino acid at the ith position (gaps excluded), and n^i is the number of times the wild-type amino acid occurs at the ith position. The Wu-Kabat coefficient of

variability considers the number of different types of amino acids that substitute the wild-type residue along with the frequency of occurrences of the wild-type amino acid residue. Therefore, if the wild-type amino acid is not well conserved, for example, if it occurs less than 50%, and the remaining 19 amino acids are found to substitute the wild-type, then Wu-Kabat's coefficient of variability takes a large number.

The sequences for all the drug targets used in this work were imported from the UniProt database[30]. For HIV-1 protease, the UniProt accession number is Q9Q288, and for human anaplastic lymphoma kinase (ALK), it is Q9UM73 (residues 1093-1400; the kinase domain on which the study has been carried out).

The ΔFV score (*F*luctuation and *V*ariability) is an empirical score that equals the product of the amino acid sequence variability and the difference in atomic fluctuation (ΔRMSF). ΔFV_1 (**Eq 2a**) is obtained by taking the difference of the atomic fluctuations for the amino acid at the ith position in the *apo*protein and the amino acid in the *holo*protein bound with either the substrate or the drug. A second fluctuation score ΔFV_2 (**Eq 2b**) is also computed by taking the difference of atomic fluctuation for the amino acid at the ith position in the substrate-bound protein and the drug-bound protein. The difference in the ΔFV_1 and ΔFV_2 scores is subtle: on the face of it, both scores identify the impact of the binding of a small molecule on the protein mobility weighted by the positional variability of an amino acid. However, the latter score identifies amino acid positions that show more pronounced interactions with the drug than the substrate, and therefore, are more likely to mutate with minimal effect on the protein function[13] (*vide infra*).

$$\Delta FV_1^i = \Delta RMSF^i_{(apo-drug\ or\ substrate)} * (Wu - Kabat\ variability\ coefficient)^i \quad \textbf{Eq 2a}$$

$$\Delta FV_2^i = \Delta RMSF^i_{(substrate-drug)} * (Wu - Kabat\ variability\ coefficient)^i$$

Eq 2b

where $\Delta RMSF^i$ is the difference in atomic fluctuations at the ith position in the amino acid sequence.

CONCLUSIONS

This work describes the development of an empirical score (ΔFV) to identify mutational hotspots. The method combines evolutionary information with binding information at the residue level(10,11). However, to the best of our knowledge, this is the first method in the structure-based class of methods that do not use free-energy changes as a measure to identify important amino acid residues involved in drug binding. Moreover, the method makes use of the best-in-class method i.e., the hidden Markov models (HMM), which is based on multiple sequence alignment, to compute the amino acid variability at each position. It has been well established that mutations generally occur at positions with low sequence conservation[10]. Such positions have been identified using Wu-Kabat's coefficient of variability. The change in the atomic fluctuations, when different ligands (substrate or drug) occupy the binding pocket, is used to identify crucial residues that are involved in the binding of respective ligands. The residues that interact more prominently with the drug than the substrate are prone to mutate with minimal or no effect on substrate binding, and hence, no compromise in the protein function is expected. Moreover, this information can also be used to design molecules which specifically show strong interaction with residues that are involved in the substrate binding, leading to a lower possibility of mutations, and thus, resistance can be delayed significantly ("substrate envelope hypothesis").

The ΔFV score developed here is the product of Wu-Kabat's coefficient of variability and atomic fluctuations. Using this score, mutational hotspots have been identified for two FDA-approved drugs against HIV-1 protease, namely nelfinavir and darunavir. This method is promising as it could identify almost all mutations that cause major resistance to darunavir and nelfinavir. Further, this method was tested on crizotinib which inhibits human anaplastic lymphoma kinase, which is a difficult target to study, as most of the mutations show complicated mechanisms of drug resistance. In this case, the method could identify 3 positions as mutational hotspots as most mutations either occur outside the active site region or enhance the catalytic efficiency without affecting crizotinib binding. However, we believe that this method has the potential to identify mutational hotspots in the active site region and suggest mutations that have not been observed as yet, in the clinic. This approach can act as a primary screen owing to its computational simplicity and speed, compared to the structure-based methods that employ free-energy computations to identify mutational hotspots. As a primary screen, this approach can help narrow down the choices of positions and amino acid substitutions to be considered for computationally demanding methods like free-energy perturbation or thermodynamic integration. To choose, the most plausible substitution, it is suggested that the multiple sequence alignment (**Figure S5**) can be employed to select the substitution for the wild-type residue. In its present dispensation, the method, however, is unable to identify mutations outside the active site region and at positions which also affect substrate binding. Also, this method is unable to identify double and triple mutations, and further developments are underway to overcome these limitations.

ASSOCIATED CONTENT

Contents of Supplementary information

Supplementary notes 1 and 2 provide explanations to supplementary **Tables S3A** and **S3B** and **S4A** and **S4B** respectively.

Table S1: List of X-ray crystal structures along with their accession codes.

Table S2: Classification of amino acid variability.

Table S3A: Drug-resistant mutations reported in the HIVdb for various inhibitors of HIV-1 protease.

Table S3B: Number and type of mutations for protease inhibitors considered in this study.

Table S4A: Backbone atoms that lie within 5 Å of the ligands as seen in the X-ray crystal structures of HIV-1 protease.

Table S4B: Side-chain atoms that lie within 5 Å of the ligands as seen in the X-ray crystal structures of HIV-1 protease.

Table S5A: $\Delta RMSF_{(apo–drug)}$ and $\Delta RMSF_{(substrate-drug)}$ values computed using side-chain atoms for nelfinavir for non-resistant positions.

Table S5B: $\Delta RMSF_{(apo–drug)}$ and $\Delta RMSF_{(substrate-drug)}$ values computed using side-chain atoms for darunavir for non-resistant positions.

Figure S1: Cα RMSD and normalized distribution plots for wild-type HIV-1 protease system. A) *Apo* HIV-1 protease, B) Substrate-bound HIV-1 protease, C) Nelfinavir-bound HIV-1 protease, D) Darunavir-bound HIV-1 protease and E) Normalized distribution plot for Cα RMSD. Runs 1, 2 and 3 indicate three independent MD runs.

Figure S2: Cα RMSD and normalized distribution plots for human anaplastic lymphoma kinase (ALK) system. A) *Apo* ALK, B) Substrate-bound ALK, C) Crizotinib-bound ALK, and D) Normalized distribution plot for Cα RMSD. Runs 1, 2 and 3 indicate three independent MD runs.

Figure S3: RMS average correlation (RAC) for HIV-1 protease system **A)** *apo*protein **B)** substrate-bound protein **C)** protease bound with nelfinavir and **D)** protease bound with darunavir.

Figure S4: RMS average correlation (RAC) for human anaplastic lymphoma kinase; **A)** *apo*protein **B)** substrate-bound protein and **C)** Crizotinib-bound protein.

Figure S5: Pie diagrams depicting the amino acid variability generated using HMM for HIV-1 protease system for positions showing mutations for nelfinavir and darunavir.

AUTHOR INFORMATION

CORRESPONDING AUTHOR

Elvis A. F. Martis, Molecular Simulations Group, Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz [E], Mumbai 400098, India. Phone: +91-7506161232. E-mail: elvis.martis@bcp.edu.in

AUTHOR CONTRIBUTIONS

EAF Martis designed and conducted computational studies, EAF Martis and EC Coutinho analysed the data, and EC Coutinho supervised the research and provided computing resources. All authors wrote, reviewed, and approved the draft of the manuscript.

FUNDING SOURCES

This work was supported by a grant from BASF SE, Ludwigshafen am Rhein, Germany to EC Coutinho which includes doctoral funding to EAF Martis.

ACKNOWLEDGEMENTS

EAF Martis and EC Coutinho are grateful to the Department of Science and Technology (DST), Department of Biotechnology (DBT) and Council of Scientific and Industrial Research (CSIR) for their financial support to build the High-Performance Computing system at the Department of Pharmaceutical Chemistry, Bombay College of Pharmacy. EAF Martis and EC Coutinho are recipients of a hardware support grant from NVIDIA Corporation. EAF Martis is indebted to BASF, Ludwigshafen, Germany for a PhD fellowship and the MCBR4 (2015) consortium (Prof. Dr P. Comba, University of Heidelberg; Prof. Dr H. Zipse LMU, Munich and Prof. Dr G. N. Sastry, IICT, Hyderabad for MCBR visiting fellowship to the Heine-Heinrich University of

Düsseldorf, Germany under Prof Dr Holger Gohlke). EAF Martis and EC Coutinho are grateful to Dr. Ian R Craig, computational chemist, BASF, Ludwigshafen, Germany for carefully reading the draft of this manuscript and providing valuable insights that significantly improved the work described in this work.

References

(1) Gulnik, S. V.; Suvorov, L. I.; Liu, B.; Yu, B.; Anderson, B.; Mitsuya, H.; Erickson, J. W. Kinetic Characterization and Cross-Resistance Patterns of HIV-1 Protease Mutants Selected under Drug Pressure. *Biochemistry* **1995**, *34* (29), 9282–9287.

(2) Schliekelman, P.; Garner, C.; Slatkin, M. Natural Selection and Resistance to HIV. *Nature* **2001**, *411* (6837), 545.

(3) Toprak, E.; Veres, A.; Michel, J.-B.; Chait, R.; Hartl, D. L.; Kishony, R. Evolutionary Paths to Antibiotic Resistance under Dynamically Sustained Drug Selection. *Nat. Genet.* **2012**, *44* (1), 101.

(4) Walsh, C. Molecular Mechanisms That Confer Antibacterial Drug Resistance. *Nature* **2000**, *406* (6797), 775.

(5) Longley, D.; Johnston, P. Molecular Mechanisms of Drug Resistance. *J. Pathol. J. Pathol. Soc. G. B. Irel.* **2005**, *205* (2), 275–292.

(6) Andersson, D. I.; Levin, B. R. The Biological Cost of Antibiotic Resistance. *Curr. Opin. Microbiol.* **1999**, *2* (5), 489–493.

(7) Gagneux, S.; Long, C. D.; Small, P. M.; Van, T.; Schoolnik, G. K.; Bohannan, B. J. The Competitive Cost of Antibiotic Resistance in Mycobacterium Tuberculosis. *Science* **2006**, *312* (5782), 1944–1946.

(8) Chellappan, S.; Kairys, V.; Fernandes, M. X.; Schiffer, C.; Gilson, M. K. Evaluation of the Substrate Envelope Hypothesis for Inhibitors of HIV-1 Protease. *Proteins Struct. Funct. Bioinforma.* **2007**, *68* (2), 561–567.

(9) Nalam, M. N.; Ali, A.; Altman, M. D.; Reddy, G. K. K.; Chellappan, S.; Kairys, V.; Özen, A.; Cao, H.; Gilson, M. K.; Tidor, B. Evaluating the Substrate-Envelope Hypothesis: Structural Analysis of Novel HIV-1 Protease Inhibitors Designed to Be Robust against Drug Resistance. *J. Virol.* **2010**, *84* (10), 5368–5378.

(10) Özen, A.; Haliloğlu, T.; Schiffer, C. A. Dynamics of Preferential Substrate Recognition in HIV-1 Protease: Redefining the Substrate Envelope. *J. Mol. Biol.* **2011**, *410* (4), 726–744.

(11) Martis, E. A.; Coutinho, E. C. Free Energy-Based Methods to Understand Drug Resistance Mutations. In *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process*; Springer, 2019; pp 1–24.

(12) Cao, Z. W.; Han, L. Y.; Zheng, C. J.; Ji, Z. L.; Chen, X.; Lin, H. H.; Chen, Y. Z. Computer Prediction of Drug Resistance Mutations in Proteins. *Drug Discov. Today* **2005**, *10* (7), 521–529.

(13) Wang, W.; Kollman, P. A. Computational Study of Protein Specificity: The Molecular Basis of HIV-1 Protease Drug Resistance. *Proc. Natl. Acad. Sci.* **2001**, *98* (26), 14937–14942.

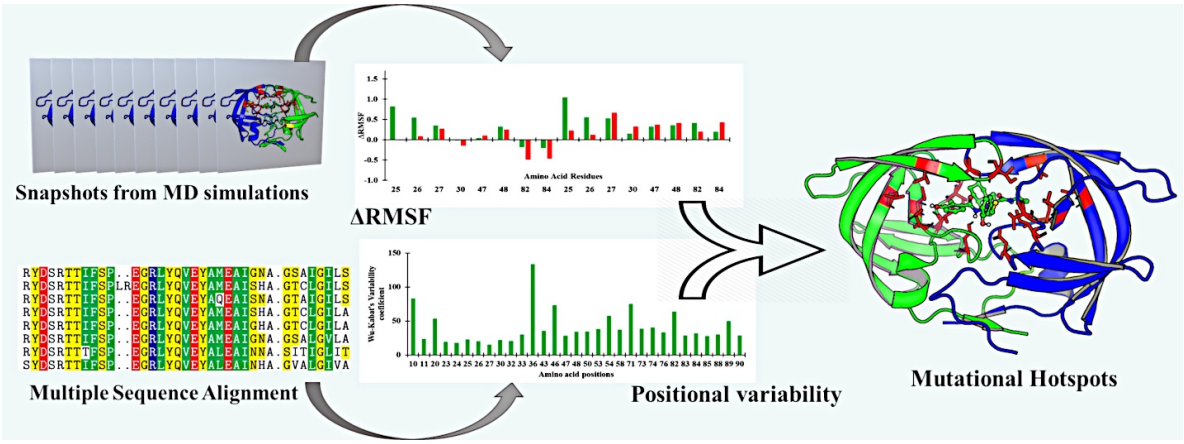
(14) Wang, W.; Lim, W. A.; Jakalian, A.; Wang, J.; Wang, J.; Luo, R.; Bayly, C. I.; Kollman, P. A. An Analysis of the Interactions between the Sem– 5 SH3 Domain and Its Ligands Using Molecular Dynamics, Free Energy Calculations, and Sequence Analysis. *J. Am. Chem. Soc.* **2001**, *123* (17), 3986–3994.

(15) Frisch, M.; Trucks, G.; Schlegel, Hb.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montgomery Jr, J.; Vreven, T.; Kudin, K.; Burant, Jc. *Gaussian 03, Revision c. 02; Gaussian*; 2004.

(16) Besler, B. H.; Merz Jr, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11* (4), 431–439.

(17) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260.
(18) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.
(19) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* **2003**, *24* (9), 1016–1025.
(20) Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *J. Chem. Theory Comput.* **2014**, *10* (3), 1228–1241.
(21) Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C. C.; Pettersen, E. F.; Goddard, T. D.; Meng, E. C.; Sali, A.; Ferrin, T. E. UCSF Chimera, MODELLER, and IMP: An Integrated Modeling System. *J. Struct. Biol.* **2012**, *179* (3), 269–278.
(22) Jorgensen, W. L.; Jenson, C. Temperature Dependence of TIP3P, SPC, and TIP4P Water from NPT Monte Carlo Simulations: Seeking Temperatures of Maximum Density. *J. Comput. Chem.* **1998**, *19* (10), 1179–1186.
(23) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N· Log (N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.
(24) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.
(25) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.
(26) Roe, D. R.; Cheatham III, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.
(27) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* **2012**, *9* (2), 173.
(28) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res.* **2005**, *33* (suppl_1), D501–D504.
(29) Kabat, E.; Wu, T. T.; Bilofsky, H. Unusual Distributions of Amino Acids in Complementarity Determining (Hypervariable) Segments of Heavy and Light Chains of Immunoglobulins and Their Possible Roles in Specificity of Antibody-Combining Sites. *J. Biol. Chem.* **1977**, *252* (19), 6609–6616.
(30) UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2014**, *43* (D1), D204–D212.
(31) Nalam, M. N.; Ali, A.; Altman, M. D.; Reddy, G. K. K.; Chellappan, S.; Kairys, V.; Özen, A.; Cao, H.; Gilson, M. K.; Tidor, B. Evaluating the Substrate-Envelope Hypothesis: Structural Analysis of Novel HIV-1 Protease Inhibitors Designed to Be Robust against Drug Resistance. *J. Virol.* **2010**, *84* (10), 5368–5378.
(32) Chellappan, S.; Kiran Kumar Reddy, G.; Ali, A.; Nalam, M. N.; Anjum, S. G.; Cao, H.; Kairys, V.; Fernandes, M. X.; Altman, M. D.; Tidor, B. Design of Mutation-resistant HIV Protease Inhibitors with the Substrate Envelope Hypothesis. *Chem. Biol. Drug Des.* **2007**, *69* (5), 298–313.

(33) Chellappan, S.; Kairys, V.; Fernandes, M. X.; Schiffer, C.; Gilson, M. K. Evaluation of the Substrate Envelope Hypothesis for Inhibitors of HIV-1 Protease. *Proteins Struct. Funct. Bioinforma.* **2007**, *68* (2), 561–567.
(34) Galindo-Murillo, R.; Roe, D. R.; Cheatham III, T. E. On the Absence of Intrahelical DNA Dynamics on the Ms to Ms Timescale. *Nat. Commun.* **2014**, *5*, 5152.
(35) Galindo-Murillo, R.; Roe, D. R.; Cheatham III, T. E. Convergence and Reproducibility in Molecular Dynamics Simulations of the DNA Duplex d (GCACGAACGAACGAACGC). *Biochim. Biophys. Acta BBA-Gen. Subj.* **2015**, *1850* (5), 1041–1058.
(36) Chang, M. W.; Torbett, B. E. Accessory Mutations Maintain Stability in Drug-Resistant HIV-1 Protease. *J. Mol. Biol.* **2011**, *410* (4), 756–760.
(37) Shafer, R. W. Rationale and Uses of a Public HIV Drug-resistance Database. *J. Infect. Dis.* **2006**, *194* (Supplement_1), S51–S58.
(38) Schock, H. B.; Garsky, V. M.; Kuo, L. C. Mutational Anatomy of an HIV-1 Protease Variant Conferring Cross-Resistance to Protease Inhibitors in Clinical Trials COMPENSATORY MODULATIONS OF BINDING AND ACTIVITY. *J. Biol. Chem.* **1996**, *271* (50), 31957–31963.
(39) Henderson, G. J.; Lee, S.-K.; Irlbeck, D. M.; Harris, J.; Kline, M.; Pollom, E.; Parkin, N.; Swanstrom, R. Interplay between Single Resistance-Associated Mutations in the HIV-1 Protease and Viral Infectivity, Protease Activity, and Inhibitor Sensitivity. *Antimicrob. Agents Chemother.* **2012**, *56* (2), 623–633.
(40) Pazhanisamy, S.; Stuver, C. M.; Cullinan, A. B.; Margolin, N.; Rao, B.; Livingston, D. J. Kinetic Characterization of Human Immunodeficiency Virus Type-1 Protease-Resistant Variants. *J. Biol. Chem.* **1996**, *271* (30), 17979–17985.
(41) Piana, S.; Carloni, P.; Rothlisberger, U. Drug Resistance in HIV-1 Protease: Flexibility-assisted Mechanism of Compensatory Mutations. *Protein Sci.* **2002**, *11* (10), 2393–2402.
(42) Hong, L.; Zhang, X. C.; Hartsuck, J. A.; Tang, J. Crystal Structure of an in Vivo HIV-1 Protease Mutant in Complex with Saquinavir: Insights into the Mechanisms of Drug Resistance. *Protein Sci.* **2000**, *9* (10), 1898–1904.
(43) Mitsuya, Y.; Winters, M. A.; Fessel, W. J.; Rhee, S.-Y.; Hurley, L.; Horberg, M.; Schiffer, C. A.; Zolopa, A. R.; Shafer, R. W. N88D Facilitates the Co-Occurrence of D30N and L90M and the Development of Multidrug Resistance in HIV Type 1 Protease Following Nelfinavir Treatment Failure. *AIDS Res. Hum. Retroviruses* **2006**, *22* (12), 1300–1305.
(44) Yedidi, R. S.; Maeda, K.; Fyvie, W. S.; Steffey, M.; Davis, D. A.; Palmer, I.; Aoki, M.; Kaufman, J. D.; Stahl, S. J.; Garimella, H. P2′ Benzene Carboxylic Acid Moiety Is Associated with Decrease in Cellular Uptake: Evaluation of Novel Nonpeptidic HIV-1 Protease Inhibitors Containing P2 Bis-Tetrahydrofuran Moiety. *Antimicrob. Agents Chemother.* **2013**, *57* (10), 4920–4927.
(45) Kaldor, S. W.; Kalish, V. J.; Davies, J. F.; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K. Viracept (Nelfinavir Mesylate, AG1343): A Potent, Orally Bioavailable Inhibitor of HIV-1 Protease. *J. Med. Chem.* **1997**, *40* (24), 3979–3985.
(46) Sun, H.; Li, Y.; Li, D.; Hou, T. Insight into Crizotinib Resistance Mechanisms Caused by Three Mutations in ALK Tyrosine Kinase Using Free Energy Calculation Approaches. *J. Chem. Inf. Model.* **2013**, *53* (9), 2376–2389.
(47) Hallberg, B.; Palmer, R. H. Mechanistic Insight into ALK Receptor Tyrosine Kinase in Human Cancer Biology. *Nat. Rev. Cancer* **2013**, *13* (10), 685.

(48) Shaw, A. T.; Friboulet, L.; Leshchiner, I.; Gainor, J. F.; Bergqvist, S.; Brooun, A.; Burke, B. J.; Deng, Y.-L.; Liu, W.; Dardaei, L. Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. *N. Engl. J. Med.* **2016**, *374* (1), 54–61.

Graphical Abstract