Coupled Solvent Dynamics and Protein Dynamics Help Drive Functional Differences in Exon-19 Deletion Mutants in the Epidermal Growth Factor Receptor (EGFR/ErbB1/HER1) Kinase Domain

Keshav Patil[a,b], Kumar D. Ashtekar [c,d], Yuko Tsutsui [c,d], Debdas Dhabal[e], Krishna Suresh [f], Mark A. Lemmon [c,d], Ravi Radhakrishnan [a,b,f,g,*]

[a] Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA
[b] Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA
[c] Department of Pharmacology, Yale University, New Haven, CT
[d] Cancer Biology Institute, Yale University, West Haven, CT
[e] Department of Chemistry, Indian Institute of Technology, Guwahati, Assam 781039, India.
[f] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[g] Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia PA

[*] Email correspondence to rradhak@seas.upenn.edu

Abstract

Deletions in Exon-19 of the epidermal growth factor receptor (EGFR) play a pivotal role in the pathogenesis of non-small cell lung cancer (NSCLC), influencing patient response to tyrosine kinase inhibitors (TKIs). Although these mutations are known to affect treatment efficacy, the precise molecular mechanisms have been unclear. Building upon recent insights from the study [DOI: 10.1038/s41467-022-34398-z], which identified two distinct mutation profiles associated with differential drug sensitivity and clinical outcomes, our research delves into the molecular dynamics that drive these variances.

We employed molecular dynamics simulations, enhanced sampling methods, and machine learning to classify Exon-19 deletion mutations into two profiles based on their conformational dynamics. Profile 1 mutations display only localized motions in key subdomains in their fluctuations about the equilibrium state, and a high affinity for ATP and consequent resistance to TKIs, while profile 2 mutations show reduced ATP binding affinity due to delocalized motion characterized by an increased flexibility between the N- and C-lobes of the EGFR kinase domain. This structural flexibility perturbs the ATP binding site, leading to decreased affinity and, heightened sensitivity to TKIs.

Our use of the INDUS technique has shed light on the collective solvent dynamics, further elucidating the coupling between long timescale solvent fluctuations and protein conformational dynamics, that likely contributes to the observations in HDX-MS studies. Our free energy analysis, covering timescales relevant to both HDX-MS and ligand interaction, provides a deeper understanding of the relationship between protein and solvent dynamics and their collective impact on drug efficacy in NSCLC with EGFR Exon-19 deletions.

Key Words: EGFR Kinase, Molecular Dynamics, Umbrella Sampling, Tyrosine Kinase Inhibitors

Introduction

Tyrosine kinase inhibitors (TKIs) are the foremost treatment option for EGFR mutations [1-3] in non-small cell lung cancer (NSCLC). Exon-19 and Exon-21 mutations represent the majority, over 90%, of the oncogenic EGFR alterations [4-6]. The most common Exon 19 mutation, constituting 33% of EGFR mutations in NSCLC, is a deletion of five amino acids (delE746-A750) in the loop that joins the third β-strand of the EGFR tyrosine kinase domain (TKD) with the vital αC helix. However, many other rarer Exon 19 deletions are seen in NSCLC patients, and appear to be associated with reduced sensitivity to EGFR TKIs [7, 8]. These EGFR Exon 19 deletion mutations differ in the length of their deletion lengths, and many are indels. Recent studies indicate that the differential drug sensitivity of Exon 19 variants to various EGFR inhibitors reflects their conformational dynamics [9]. Comparison of a set of Exon 19 deleted variants suggested that their drug binding affinity (with erlotinib) was essentially unchanged, but they showed alter K_M values for ATP, suggesting that variation in ATP binding affinity is the main determinant of (ATP-competitive) drug sensitivity. Grouping the variants studied according to their ATP binding affinity and analyzing their structural dynamics through HDX-MS to acquire exchange profiles allowed two distinct groups or profiles to be identified. Profile 1 mutations (e.g., L747P, del747-750-insP) showed higher ATP binding affinity, translating to a lower K_M, whereas Profile 2 mutations (e.g., del746-750, del-747-753-insS) exhibited increased flexibility and disorder in the ATP binding site, resulting in reduced ATP binding affinity. This variable ATP binding affinity contributed to drug resistance due to the heightened competition between the kinase inhibitor and ATP for the binding site. Kaplan-Meier survival analyses further indicated that patients with Profile 1 mutations had a shorter median progression-free survival on erlotinib treatment compared to those with Profile 2 mutations.

Our aim in this paper is to use molecular dynamics (MD) to delve into the molecular underpinnings of the differential dynamics in exon 19 deletion mutants. We attribute the differential TKI sensitivity of Exon 19-mutated variants to altered protein conformational dynamics, backbone hydrogen bond networks, solvent accessibility, and electrostatic interactions. We propose using MD as a cohesive framework to decode both the hydrogen deuterium exchange profiles and the structural origins of this exchange, as a critical step in comprehending how mutations affect protein function, particularly in kinases. Protein dynamics are crucial to their function and biomolecular interactions. Hydrogen-deuterium exchange mass spectrometry (HDX-MS) is an effective method for probing protein dynamics. The well-established two-step model of HDX, proposed by Linderstrom-Lang [10], is utilized to interpret the protein dynamics observed, where the protein's amide N-H hydrogen transitions from a closed to an open state before exchanging with solvent deuterium. HDX-MS, in conjunction with mass spectrometry, is particularly advantageous for studying larger protein complexes without size restrictions, enhancing the molecular understanding of the dynamics involved in kinases and the impact of mutations on these dynamics. HDX provides insights into protein dynamics through changes in hydrogen bond networks, and some conformational changes can be 'HDX-silent' if they cause only minimal changes in hydrogen bonding. For instance, the rupture of Heme-Fe(III) contact in cytochrome c causes significant reorientation of a loop with little change in hydrogen bonding, and is HDX-silent [11]. Correlating HDX-MS data with MD simulations has the potential to be of great value is for understanding the molecular origins behind the observed deuterium exchange of biomolecules studied in HDX [12, 13].

Whereas MD simulations typically focus on the microsecond scale, HDX operates in much larger scales from seconds to hours. Accordingly, HDX can capture protein dynamics across a vast range of timescales, encompassing protein folding/unfolding, conformational transition states, ligand binding, and other biomolecular interactions. To use MD to link HDX results to their molecular origins, therefore, the timescale of MD must be expanded, which can be achieved using enhanced sampling techniques. However, implementing such techniques requires prior knowledge of the collective variables that govern protein transitions, which can be difficult to predict. To extend MD analysis in the context of HDX-MS, we apply Indirect Umbrella Sampling (INDUS), a technique commonly used for studying protein hydration dynamics [14]. This allows us to simulate the coupling between the collective solvent fluctuations and protein conformational dynamics to gain insight into long-timescale fluctuations in protein conformations in the EGFR Exon-19 deletion mutant systems. In this report, we combine MD simulations, machine learning, and HDX-MS to elucidate the impact of EGFR Exon-19 deletion mutations on short timescale (microsecond) kinase conformational dynamics. We demonstrate the critical role of hydrogen bonding networks in shaping these dynamics and, through the use of INDUS, reveal the intricate relationship between collective solvent dynamics and long timescale (i.e., milliseconds and larger) conformational fluctuations, thus improving the integration of MD simulations with HDX-MS data.

Results

Alongside their analysis of TKI sensitivity of exon 19-deleted variants of the measured purified EGFR tyrosine kinase domain (TKD), van Alderwerelt van Rosenburgh *et al*. used HDX-MS to identify conformational dynamics as a possible origin for the differences, and proposed a basis for classifying uncommon exon 19 variants that may have predictive clinical value [9]. They showed that native state dynamics of exon 19 variants differ for profile 1 and profile 2 variants as shown in Fig. S1C. Profile 1 (more TKI-resistant) variants (e.g. ΔL747-A750InsP and L747P) showed generally lower degrees of exchange in the ATP-binding region of the TKD than profile 2 (more TKI-sensitive) variants (e.g. ΔE746-A750 and ΔL747-P753InsS). How the altered conformational dynamics as indicated by HDX translates to altered ATP binding affinity, however, is unclear. We chose variants in profile 1 and 2 that differ by only one amino acid to analyze by molecular dynamics in order to investigate the origin of this difference in conformational dynamics,

Principal dynamics and cross correlations at the subdomain level correctly profile EGFR variants

We performed 4 μs MD simulations of all profile 1 and profile 2 variants studied experimentally by van Alderwerelt van Rosenburgh *et al*. as described in Methods. We first compared the MD trajectories for the profile 1 (resistant) Δ747-750InsP and profile 2 (sensitive) Δ746-750 systems. We performed principal component analysis (PCA) of the two MD trajectories and projected the configurations visited in the respective trajectories along the top two principal components (Fig. 1A). Based on these two-dimensional projections, we identified clusters according to the relative frequency of conformations visited; a stark difference in the PCA projections is evident. The difference can also be visualized by comparing the structures corresponding to the centroids of each of the clusters as shown in Fig. 1B.

We summarize the results of the PCA on all other variants studied experimentally in Table 1, wherein the principal motions in certain kinase subdomains are listed. We also look into the essential motions in N-lobe and C-lobe, since their motion provides global flexibility for the ATP binding pocket. We observe that the motions in the profile 1 variants occur largely around P-loop, αC-helix and Activation loop, whereas these motions also extend to N–lobe and C-lobe regions for the profile 2 variants. To provide a quantitative description and classification of these principal motions we next computed residue-wise dynamical cross correlation matrices from the atomic fluctuations, and coarse-grained these matrices to subdomain-level cross correlations as described in Methods and Table 2.

The coarse-grained correlation matrices for profile 1 variants (and wild-type) are shown in Fig. 2A, and those for profile 2 variants are shown in Fig. 2B. The qualitative categories of low, medium, and high degrees of cross correlations summarized in Table 1 are quantitatively borne out in Fig. 2. The most immediately obvious difference between profile 1 and profile 2 variants in these matrices is the greater delocalization of cross-correlations in profile 2 mutants, engaging regions in both the N- and C-lobes of the TKD. Both profiles show high degrees of correlated motion primarily in the N-lobe, and cross-correlated motions between subdomains of the N-lobe (strands β1-5, and the P-loop). Profile 1 variants show little cross-correlation outside these regions, whereas profile 2 variants, notably Δ747-750InsP and Δ746-751InsA, show significant cross correlation involving C-lobe β-strands and α-helices as well as the Activation loop. To further categorize these differences between profile 1 and profile 2 variants, we compared the dyamical cross correlations (DCCs) by performing a pairwise distance calculation and constructed a hierarchical cluster (Fig. 2C); we find that this unsupervised clustering is remarkably efficient in classifying the EGFR variant systems into the two separate profiles with excellent agreement with what was observed in the experiments (REF). The approach readily identifies the wild-type TKD as well as Δ747-749, Δ747-750InsP, and L747P as profile 1. Their experimental K_M values for ATP range from 12 to 23 μM (Ref). The Δ746-750, Δ752-759 and Δ746-751InsP cluster differently and have significantly higher values for $K_{M,\ ATP}$ (91 to 164 μM). We note that two systems Δ747-753InsS and Δ747-751InsP are misclassified in this analysis. We surmise that a possible reason for this misclassification is the limited sampling of conformations in molecular dynamics (spanning just 4 μs) in comparison to the much longer timescales (seconds-to-hours) over which exchange occurs in HDX experiments. We address this limitation through enhanced sampling discussed later.

Hydrogen bond and solvent accessibility observed in unbiased simulations explain HDX data and provide validation of the observed changes in the protein conformational dynamics

To validate the molecular origins of the observed variations in the dynamics across profile 1 and profile 2 variant, we sought to compare the MD results with the experimentally observed HDX rates. The arrangement of the hydrogen bond network is a major determinant of the observed percent exchanges in HDX, as hydrogen bonds between protein residues must be broken to enable exchange of the hydrogen with deuterium in the solvent. Solvent accessibility also impacts HDX, however, as residues deeply buried in the protein have low rates of exchange due to lack of access to contact solvent molecules. As a measure of hydrogen bond lability, which should correlate with HDX based on these arguments, we plot "1 minus the hydrogen bond occupancy" from MD simulations of the EGFR deletion mutants as described in Methods. Comparing these results in

Fig. 3A with the experimentally reported HDX results (Fig. S1C) supports the premise that hydrogen bond occupancy in MD is lower where HDX is higher.

To obtain a quantitative mapping between the computationally-determined H bond occupancy (Fig. 3A), solvent-accessible surface area (SASA) per residue (Fig. 3B), and the experimentally obtained percent HDX (Fig. S1C), we implement an information theory-based metric and a machine learning model. We use these to combine information in H bond occupancy and SASA to measure the explainability of the observed percent exchange in HDX studies. Across all systems analyzed, we find that combining SASA and Hbond occupancy leads to an increase in the mutual information, i.e., more information on exchange is gained (see Table 3).

While it is challenging to find an equation that can combine H bond occupancy and SASA to provide % HDX as an output, machine learning models are ideal tools to combine multimodal data and can be trained to provide such a mapping. We trained a decision tree to take as input the H bond occupancy and SASA information as features and provide as output the % HDX (Fig 3C). Based on this decision tree regression model, we estimate the SHAP (SHapley Additive exPlanations) values [15], which measure the importance of each feature in the model in predicting its output while quantifying the nature of the contribution. Features with positive SHAP values positively impact the prediction, whereas those with negative values have a negative impact – with the magnitude of the SHAP value corresponding to the strength of the effect. As shown in Fig. 3D, H bond occupancy is more important towards predicting the % HDX than SASA, with higher H bond occupancy values (red) being associated with negative SHAP values; i.e. reducing the % HDX. Higher values for SASA (red), by contrast, are associated with positive SHAP values; i.e. with increased % HDX – consistent with our expectations and the results of Fig. 3A,B. Our results thus show that the observed dynamics captured in terms of principal motions are reflected in the observed exchange profiles in HDX experiments, thereby validating the computational findings.

Conformational fluctuations coupled to solvation effects at long timescale explain the distinction between some of the EGFR Exon-19 deletion variants

Although we have investigated the molecular origins behind the observed differences in HDX for the different EGFR Exon 19-deleted variants based on the trajectories of limited timescale simulations, we note they are still not sufficient to fully explain the misclassified variants in the hierarchical cluster shown in Fig. 2C. In Fig. 4A, we observe that when % HDX is plotted against H bond occupancy for the wild type EGFR TKD, certain sequences fall in the lower left quadrant where both the H bond occupancy and % HDX simultaneously have low values (See Appendix). While this trend is counterintuitive, we note these same peptides also have low SASA (marked in red), which explains their unexpectedly low % HDX.

In the case of the Δ747-753InsS system, one of the outliers in Fig. 2C, we find that certain sequences fall in under the upper right quadrant when % HDX is plotted against H bond occupancy (see Fig. 4B and Appendix), indicating high H bond occupancy and higher % HDX. This is trend is inconsistent with our intuition and is also not explained by the SASA values. We hypothesize that this is the result of limited timescale dynamics information provided by unbiased molecular dynamics, wherein the conformational sampling does not adequately represent partial unfolding

of the backbone that could occur at timescales longer than the few microseconds captured by MD, but still within the experimental observation time of HDX experiments.

With this in mind, we computed the free energy change associated with solvent density fluctuations using Indirect Umbrella Sampling (INDUS) as depicted in Fig. 4C. In HDX studies, the protein is constantly undergoing transitions between partially folded and unfolded states, which is generally coupled with fluctuations in the density of water molecules in the probe volume around the protein. This is exactly the thermodynamic landscape explored in INDUS as depicted in Fig. 4D, where we compare the structures of the equilibrium conformation with that corresponding to the free energy of 30 k_BT. We structurally aligned the average structure from INDUS (from the zone of 30 k_BT) with respect to the equilibrium structure, and find the RMSD to be 4.51 Å, suggesting only local changes in the secondary structure, and that the protein is not completely unfolding. The choice of this region of the free energy (~30 k_BT) was based on the barrier that can be overcome in the HDX timescale of the order of seconds. We found that this mis-classified profile 2 system (Δ747-753InsS) was more elastic with respect to movement of water in and out of the system compared with the other deletion variants, since the free energies associated with water density fluctuations were the lowest of all the EGFR deletion variants (Fig. 4C).

We can further explain the outlier results in Fig. 4B, wherein certain peptides that showed higher H bond occupancies also showed high % HDX. By extracting structures from the activated state of ~30 k_BT (gray box marked in Fig. 4C), we can compare the H bond profiles for the equilibrium structures with those corresponding to the activated state. We note that in the activated state, the Δ747-753InsS system undergoes a loss in H bond occupancies in the outlier sequences. Fig. S3 depicts the difference in the H bond occupancy profile between the equilibrium and activated states; the outlier sequences for Fig. 4B are marked with a black bar, which show reduced H bond occupancies in the activated state but high occupancy in the equilibrium state. Taken together, these results explain the paradoxical result of high H bond occupancy coupled with high exchange rates in Fig. 4B. Moreover, they underscore the importance of long-timescale fluctuations in backbone conformations coupled to solvent density fluctuations which are important to consider in interpreting HDX studies.

Discussion and Conclusion

In this work, we elucidate the impact of EGFR Exon-19 deletion mutations on kinase conformational dynamics by integrating MD simulations, machine learning, enhanced sampling and HDX-MS. We analyze MD simulations of protein dynamics using Principal Component Analysis and record distinct modes of conformational excursions for the EGFR deletion mutants in two TKI resistance 'profiles' identified experimentally. We observe that the motions in profile 1 (resistant) variants are largely localized around the nucleotide binding P-loop, the αC helix and the Activation loop, whereas for profile 2 (sensitive) variants the motions also extend to the N–lobe and C-lobe regions. We established a quantitative metric to support this qualitative observation by defining a dynamical cross correlation metric based on a coarse-grained representation of the residues into 18 sub-domains.

When binding to a prototypical kinase such as EGFR, ATP is bound at the interface between the C-lobe and the N-lobe [16]. The ATP phosphates interact closely with the P-loop that connects

strands β1 and β2 and also interact with a conserved lysine in strand β3. The N- and C-lobes are connected by a linker sequence referred to as the hinge that interacts with the adenine base of the ATP molecule through one or two H bonds [16, 17]. Since profile 1 and profile 2 variants differ in their ATP binding affinities as estimated in $K_{M,ATP}$ measurements, we wondered if the differences in conformational dynamics between the two profiles (Fig. 1) translates into changes in ATP binding. We therefore performed molecular docking studies across all of the clusters observed in the dynamics simulations (see Fig. 5). Our results indicate that indeed the ATP binding affinity relative to the binding affinity of erlotinib is a clear differentiator of profile 1 and profile 2 mutants, with profile 1 variants (L747P, Δ747-750InsP) showing an increased relative affinity for ATP. More generally, localized motions that do not induce relative conformational fluctuations between the C-lobe and the N-lobe preserve the order of the ATP binding site as seen in prile 1 variants. Largescale delocalized motion with relative conformational fluctuations between the two lobes as seen in profile 2 variants, however, induces disorder in the ATP binding site and negatively impacts ATP binding. This effect is felt in the P-loop region as well as the adjoining β1-β3 strands as clearly evidenced by the dynamical cross correlations between these sub-domains and those in the C-lobe (Fig. 2).

By performing a pair-wise comparison of the subdomain-level dynamical cross correlations, we successfully classified the EGFR mutated systems through a hierarchical clustering analysis into profile 1 and profile 2 sets. Remarkably, all except two mutations, Δ747-753InsS and Δ747-751InsP were classified correctly with this approach. For the two misclassified mutants, we hypothesized a role for collective solvent fluctuations in shaping the overall protein dynamics in addition to the intrinsic-protein dynamics in a solvated environment, and we probed this further by analyzing our results in light of HDX experiments. Our studies also provided a machine learning-based interpretability metric, SHAP values, that quantify the contributions of hydrogen bonding and solvent accessibility in explaining the experimentally observed percentage exchanges in HDX. We conclude that, although H bond occupancy is the higher ranking metric in predicting HDX, combining it with SASA enhances prediction of the % HDX values. These observations led to the hypothesis that solvent accessibility coupled to long-timescale motion in protein conformational dynamics can play a role in exchange kinetics and could be important for our misclassification of the Δ747-753InsS and Δ747-751InsP variants. Hydration dynamics is an important component of protein dynamics but generally is a diverse timescale phenomenon and is therefore not typically well quantified when correlating MD studies with HDX. We implemented the INDUS enhanced sampling method to quantify the thermodynamic landscape associated with fluctuations in solvent density and find that the Δ747-753InsS system has the lowest free energy barrier for invasion of solvent molecules into the probe volume, which is coupled to protein conformational flexibility. Our results comparing the H bond and solvent accessibility in the ground and excited directly demonstrate that conformation fluctuations coupled with solvent dynamics can contribute to exchange, and the exchange profiles match that of the HDX experiments. We believe that integrating MD with INDUS and appending it with data science tools can be a powerful tool to improve the interpretability of HDX results and enable better mapping of molecular origins behind the observed deuterium exchanges of biomolecules studied in HDX. More importantly, such physics-based analysis involving enhanced sampling can be critical in explaining the origin of drug sensitivity and patient survival in cancer patients who present a varied genomic landscape by way of missense mutations and in-frame deletions and insertions.

Methods

Molecular Dynamics (MD) Simulations

We constructed variants of mutations in the EGFR structure (1M14 in protein data bank) using MODELLER [18]. Specifically, in addition to the wildtype, we studied L747P, ΔL747-E749, ΔL747-A750InsP, ΔE746-A750, ΔS752-I759, ΔE746-T751InsA, ΔL747-T751InsP, ΔL747-P753InsS. To investigate the molecular underpinnings of observed differences in exchange profiles, unbiased molecular dynamics (MD) simulations totaling 4 microseconds (4 μs) were conducted using the Desmond simulation package [19] within the Schrödinger suite 2020-3, operating on a CentOS 7 Linux platform. Simulations were performed in the NPT ensemble at 300 K and 1 bar pressure, employing the Optimized Potentials for Liquid Simulations (OPLS 2005) force field [20, 21]. The protein complex was centered in an orthorhombic box, and TIP3P water molecules, along with buffer atoms, were filled within a 10 Å distance from protein atoms to the box's edge. To neutralize the system, counter ions (Na+ and Cl−) were added randomly, resulting in a salt concentration of 150 mM. Long-range electrostatic interactions were computed using the particle mesh Ewald method [22, 23], with a Coulomb interactions cutoff radius of 9 Å. Non-bonded forces were integrated using the r-RESPA integrator [24]. Temperature and pressure control were achieved using the Nose Hoover chain coupling with a 1.0 ps relaxation time at 300 K, and the Martyna-Tobias-Klein chain isotropic coupling barostat method [25] with a 2.0 ps coupling constant for pressure control.

Data Analysis

Trajectory analyses were performed, unless otherwise noted, on the whole trajectory using the MDAnalysis package [26] in Python. Python scripts that implement these analyses are available on GitHub [27].

Trajectories were saved at intervals of 100/200 ps for subsequent analysis, including examination of polar interactions and root mean square deviations (RMSD) to monitor simulation stability. Principal Component Analysis (PCA) was performed on the trajectories using the GROMACS functionality gmx anaeig, projecting the results onto the top two components. Hydrogen bond computation followed specified criteria, and solvent accessible surface area calculations were conducted using GROMACS functionality gmx sasa.

<u>Dynamical Cross Correlation (DCC):</u> The normalized covariance matrix gives the traditional residue-level dynamical cross correlation (DCC). given by:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\langle \Delta r_i^2 \rangle^{1/2} \langle \Delta r_j^2 \rangle^{1/2}}$$

here, $\Delta r_{i,k} = r_{i,k} - \mu_{r,i}$, where $r_{i,k}$is coordinate of residue i in frame k; and $\mu_{r,i}$ is the mean location of residue i over total N frames. <u>Coarse-grained DCCs</u>: to obtain correlations at the sub-domains or secondary structure level and ascribe a broad mechanism, we coarse-grained the residue-level DCC matrix to their secondary structure [28] as follows. The following residue number ranges were assigned to each of the 18 subdomains (see Table 1). The NxN BW-DC matrix

was reduced to a 18×18 coarse-grained matrix by averaging the correlation coefficients in each sub-block.

The H bond occupancy per residue was computed based on the number of frames a residue has a hydrogen bond / total number of frames. The H bond occupancy of the peptide was obtained using the mean of H bond occupancy over the residues of the peptides. The SASA per residue was computed using gmx_sasa computing the sasa for N and H of each residue and the SASA per peptide was taken as the mean of this SASA per residue for the residues in a peptide.

From information theory, the mutual information between two variables is defined as I(Z; X) = H(Z) - H(Z|X) [29]. I(Z) represents the reduction in the uncertainty of Z due to information on X; in our implementation, Z = % exchange in HDX and X is the H bond occupancy. H(Z) is defined as H(Z) = - P(Z)log(P(Z)), where P(Z) is the distribution of Z. Similarly, joint mutual information is defined I(X, Y; Z), where Z= % exchange and X= H bond occupancy and Y is the SASA and I(X, Y; Z) is defined as the information that is gained in Z by knowing X and Y. We compute mutual information, joint mutual information and the normalized mutual information and tabulate them for EGFR deletion systems (Table 3).

Indirect Umbrella Sampling (INDUS) Simulations

We compute the free energy change associated with solvent density fluctuations using the Indirect Umbrella Sampling (INDUS) [14] where the variable that is biased is the number of water molecules in the probe volume of the protein. INDUS simulations, an enhanced sampling technique, were performed using GROMACS 4.5.3 patched with INDUS, with the probe volume defined around the protein aggregating probe volumes (radius 0.6 nm) over heavy atoms. The obtained free energy from INDUS simulations was unbiased using the weighted histogram analysis method. Trajectories were processed in Python for further analysis, and EGFR structure mutations were introduced using MODELLER. All scripts and codes utilized in these analyses are publicly available on GitHub [27].

We employed the indirect umbrella sampling (INDUS) technique to investigate the free energetics of water intrusion and displacement from the solvation shell of the wild-type protein and its 4 different mutations. The INDUS method allows modulating the number of water molecules, N_v, within the solvation shell of volume v of a protein by introducing a bias on a coarse-grained water number $\widetilde{N}_v$. Unlike N_v, $\widetilde{N}_v$ represents a continuous variable of particle positions and does not induce abrupt forces upon application of biasing potential [30-35]. The observation volume v for the INDUS simulation is defined as the combination of sub-volumes with radius $r_v = 0.6$ nm, each bound to all the heavy atoms of proteins. The collective assembly of these sub-volumes approximately depicts the solvation shell of the protein. A simulation snapshot of the WT protein, highlighting the observation volume v is shown in Figure S3. It should be noted that the observation volume roughly mimics the shape of the protein and can dynamically adjust with changes in its conformation [33, 34].

A biasing potential of the form $U_b(\widetilde{N}) = \frac{k}{2}\left(\widetilde{N} - \widetilde{N}_b^*\right)^2$ with force constant k is employed for each protein, as described in Table 4. We first performed an unbiased simulation for 10 ns to compute the variance in the number of water molecules ($\langle \delta N^2 \rangle_0$) within the observation volume of a

specific protein [30-35]. The value of k is selected such that it is at least four times greater than $\frac{k_B T}{\langle \delta N^2 \rangle_0}$, where k_B is the Boltzmann constant, and T is the temperature at which the simulation is performed. The final configuration of the unbiased simulation is utilized for biased simulation, with the biasing number of water molecules ranging from 1700 to 2600 with intervals of 50 (Table 4). The duration of simulation for each biasing window is taken to be 10 ns. We used the weighted histogram analysis method (WHAM) to combine the biased probability distributions to estimate the free energy $\beta F_v(\tilde{N}) = -\ln(P_v(\tilde{N})$, where $P_v(\tilde{N})$is the probability of observing $\tilde{N}$ water molecules in v and $\beta = 1/k_B T$.

Since the different EGFR Exon-19 deletion mutant have a different numbers of water molecules in the probe volume at equilibrium (N_{eq}), we shift the plots so that the x-axis represents the deviation of the water molecules in the probe volume from the equilibrium value, N_{eq}. The y-axis is also shifted so that the minimum corresponds to zero free energy.

Molecular Docking

Preparation of protein and ligand structures*:* The ligands were prepared by adding missing atoms, assigning charges, and optimizing the geometry following standard protocol in Schrodinger's LigPrep package. The EGFR receptor structures were prepared by adding missing atoms, assigning charges, optimizing the hydrogen bonding network, and performing restrained minimization following standard protocol using Schrodinger's Protein Preparation Wizard tool. The receptor structures were then aligned to a known crystal structure of the EGFR (PDB: 1M17) tyrosine kinase domain in complex with erlotinib. The centroid of erlotinib is used as the center for the docking grid, as this the best estimate of EGFR's binding site.

Simulation*:* Schrodinger's (v. 2019.4) Glide software was used to performing docking simulations for each of the cluster identified through PCA [36, 37]. Ligands were docked using default Glide SP settings, except that "Enhanced Sampling" was set to 4, to increase the number of ligand conformers considered. For each ligand, a up to 100 of the most highly ranked poses were considered for further analysis.

Prediction of Binding Pose and Affinity using ComBind*:* The poses outputted by the Glide software package are scored using the internal GlideScore function, traditionally used as an estimate of binding affinity. Studies have shown that this internal score is actually a poor estimate of binding affinity, which can lead to inaccurate prediction of binding pose. Recently, new methods have been created, combining information from both physics-based and ligand-based approaches to provide better estimates of binding affinity. One such method, ComBind [38], additionally considers other ligands known to bind the target protein (whose binding poses are not known), resulting in more accurate predictions. The ComBind methodology is used in our study for the prediction of the binding pose and affinity of our ligands to the EGFR systems. First, a set of diverse ligands known to bind EGFR were compiled. The following is a list of these 22 "helper" ligands:

gefitinib, zorifertinib, sapitinib, poziotinib, tarlox, cln081, az5104, mobocertinib, lapatinib, icotinib, neratinib, dacomitinib, alomunertinib, olmutinib, pyrotinib, brigatinib, vandetanib, mavelertinibnazartinib, abivertinib, rociletinibnaquotinib, lazertinib, alflutinib.

The helper ligands and candidate ligands (ligands of interest) are individually docked to the target EGFR systems using the simulation protocol previously described. Combind is used to predict poses for all of the helper ligands. We omit the "substructure similarity" term in calculation of Combind scores given the diverse range of high affinity binders chosen for our set of helper ligands. For each candidate ligand, a pose is selected that minimizes the Combind score with respect to the helper ligands. The Combind score of each candidate ligand in its predicted pose is used as a prediction of its binding affinity to the target EGFR system. Of note, a more positive Combind score equates to a higher binding affinity. The pose that yielded the lowest (most negative) docking score that also preserved the known modes of binding was recorded and reported. The detailed procedure for identifying know binding modes is described in earlier works [17, 39], which was closely followed here.

Appendix

The peptide sequences that fall under the lower left quadrant of the WT percent exchange vs. Hbond occupancy (Fig 4A) :

['LRILKETEFKKIKVLGSGA',
'RILKETEF',
'VMASVDNPHVCRL',
'SVDNPHVCRL',
'LGICL',
'LTSTVQL',
'TSTVQL',
'VQLITQL',
'MPFGCL',
LDYVREHKDNIGSQYL',
'YLEDRRLVHRDLAA',
'LEDRRLVHRDLAARN',
'LEDRRLVHRDLAARNVL',
'RRLVHRDLAARNVL',
'VLVKTPQHVKITDFGL',
'VKTPQHVKITDFGLAKLLGAE',
'KEYHAEGGKVPIKWMAL',
'YHAEGGKVPIKWMAL',
'ESILHRIYTHQSDVW',
'ESILHRIYTHQSDVWS',
'HRIYTHQSDVWS',
'ISSILEKGERLPQPPICT',
'SILEKGERLPQPPICT',
'ILEKGERLPQPPIC',
'ILEKGERLPQPPICT',
'EKGERLPQPPICT',
'IDVYM',

’IMVKC’,
’IMVKCWM’,
’WMIDADSRPKFREL’,
’MIDADSRPKFRE’,
’MIDADSRPKFREL’,
’IDADSRPKFRE’,
’IDADSRPKFREL’,
’DADSRPKFRE’,
’FSKMARDPQRY’,
’MARDPQRY’,
’ARDPQRY’

The peptide sequences that fall under the upper right quadrant of the del747-753-insS percent exchange vs. Hbond occupancy (Fig 4B) :

[’DYVREHKDNIGSQ’,
’DYVREHKDNIGSQYLL’,
’IAKGMNY’,
’ISSIL’,
’ISSILEKGERLPQPPICT’,
’LVIQGDE’,
’YRALMDE’,
’YRALMDEEDM’,
’YRALMDEEDMD’

References

1. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2004. **350**(21): p. 2129-39.
2. Sordella, R., et al., *Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways.* Science, 2004. **305**(5687): p. 1163-7.
3. Sharma, S.V., et al., *Epidermal growth factor receptor mutations in lung cancer.* Nat Rev Cancer, 2007. **7**(3): p. 169-81.
4. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Research, 2014.
5. Graham, R.P., et al., *Worldwide Frequency of Commonly Detected EGFR Mutations.* Archives of Pathology & Laboratory Medicine, 2017. **142**(2): p. 163-167.
6. Roskoski, R., *Small molecule inhibitors targeting the EGFR/ErbB family of protein-tyrosine kinases in human cancers.* Pharmacological Research, 2019. **139**: p. 395-411.
7. Xu, H., et al., *Heterogeneous Response to First-Generation Tyrosine Kinase Inhibitors in Non-Small-Cell Lung Cancers with Different EGFR Exon 19 Mutations.* Targeted Oncology, 2020. **15**(3): p. 357-364.

8. Xu, C.-w., et al., *Molecular Characteristics and Clinical Outcomes of EGFR Exon 19 C-Helix Deletion in Non–Small Cell Lung Cancer and Response to EGFR TKIs.* Translational Oncology, 2020. **13**(9): p. 100791.
9. van Alderwerelt van Rosenburgh, I.K., et al., *Biochemical and structural basis for differential inhibitor sensitivity of EGFR with distinct exon 19 mutations.* Nat Commun, 2022. **13**(1): p. 6791.
10. Englander, S.W., et al., *Hydrogen exchange: The modern legacy of Linderstrøm-Lang.* Protein Science, 1997. **6**(5): p. 1101-1109.
11. Scrosati, P.M., V. Yin, and L. Konermann, *Hydrogen/Deuterium Exchange Measurements May Provide an Incomplete View of Protein Dynamics: a Case Study on Cytochrome c.* Analytical Chemistry, 2021. **93**(42): p. 14121-14129.
12. Martens, C., et al., *Integrating hydrogen–deuterium exchange mass spectrometry with molecular dynamics simulations to probe lipid-modulated conformational changes in membrane proteins.* Nature Protocols, 2019. **14**(11): p. 3183-3204.
13. Devaurs, D., D.A. Antunes, and A.J. Borysik, *Computational Modeling of Molecular Structures Guided by Hydrogen-Exchange Data.* Journal of the American Society for Mass Spectrometry, 2022. **33**(2): p. 215-237.
14. Rego, N.B., E. Xi, and A.J. Patel, *Protein Hydration Waters Are Susceptible to Unfavorable Perturbations.* Journal of the American Chemical Society, 2019.
15. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4768–4777.
16. Arter, C., et al., *Structural features of the protein kinase domain and targeted binding by small-molecule inhibitors.* J Biol Chem, 2022. **298**(8): p. 102247.
17. Liu, Y., N.J. Agrawal, and R. Radhakrishnan, *A flexible-protein molecular docking study of the binding of ruthenium complex compounds to PIM1, GSK-3beta, and CDK2/Cyclin A protein kinases.* J Mol Model, 2013. **19**(1): p. 371-82.
18. Fiser, A. and A. Sali, *Modeller: generation and refinement of homology-based protein structure models.* Methods Enzymol, 2003. **374**: p. 461-91.
19. Bowers, K.J., et al. *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*. in *SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. 2006.
20. Jorgensen, W.L. and J. Tiradorives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1657-1666.
21. Mu, Y.G., D.S. Kosov, and G. Stock, *Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments.* Journal of Physical Chemistry B, 2003. **107**(21): p. 5064-5073.
22. *Molecular Dynamics Simulations of Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins.* J. Amer. Chem. Soc., 1995. **117**: p. 4193-4194.
23. Essmann, U., *A smooth particle mesh Ewald method.* J Chem Phys, 1995. **103**: p. 8577-8593.
24. Tuckerman, M.E., B.J. Berne, and G.J. Martyna, *Reversible Multiple Time Scale Molecular Dynamics.* J. Chem. Phys., 1992. **97**: p. 1990-2001.

25. Glenn, J.M., J.T. Douglas, and L.K. Michael, *Constant pressure molecular dynamics algorithms.* The Journal of Chemical Physics, 1994. **101**(5): p. 4177-4189.
26. Gowers, R.J., et al., *MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations*, in *Conference: PROC. OF THE 15th PYTHON IN SCIENCE CONF. (SCIPY 2016) ; 2016-07-11 - 2016-07-11 ;*. 2019: United States. p. Medium: ED; Size: 98.
27. Patil, K., *GitHub Repository for EGFR Exon 19 Deletion Mutant Simulations.* 2024.
28. Mansour, S.J., et al., *Interdependent Domains Controlling the Enzymatic Activity of Mitogen-Activated Protein Kinase Kinase 1.* Biochemistry, 1996. **35**(48): p. 15529-15536.
29. Zeng, Z., et al., *A novel feature selection method considering feature interaction.* Pattern Recognition, 2015. **48**(8): p. 2656-2666.
30. Patel, A.J., et al., *Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling.* J Stat Phys, 2011. **145**(2): p. 265-275.
31. Patel, A.J., P. Varilly, and D. Chandler, *Fluctuations of water near extended hydrophobic and hydrophilic surfaces.* J Phys Chem B, 2010. **114**(4): p. 1632-7.
32. Xi, E., R.C. Remsing, and A.J. Patel, *Sparse Sampling of Water Density Fluctuations in Interfacial Environments.* J Chem Theory Comput, 2016. **12**(2): p. 706-13.
33. Rego, N.B., E. Xi, and A.J. Patel, *Protein Hydration Waters Are Susceptible to Unfavorable Perturbations.* J Am Chem Soc, 2019. **141**(5): p. 2080-2086.
34. Dhabal, D., et al., *Characterizing the Interplay between Polymer Solvation and Conformation.* J Phys Chem B, 2021. **125**(20): p. 5434-5442.
35. Jiang, Z., et al., *Characterizing Solvent Density Fluctuations in Dynamical Observation Volumes.* J Phys Chem B, 2019. **123**(7): p. 1650-1661.
36. Friesner, R.A., et al., *Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes.* Journal of Medicinal Chemistry, 2006. **49**(21): p. 6177-6196.
37. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* J Med Chem, 2004. **47**(7): p. 1739-49.
38. Paggi, J.M., et al., *Leveraging nonstructural data to predict structures and affinities of protein-ligand complexes.* Proc Natl Acad Sci U S A, 2021. **118**(51).
39. Park, J.H., et al., *Erlotinib binds both inactive and active conformations of the EGFR tyrosine kinase domain.* Biochem J, 2012. **448**(3): p. 417-23.

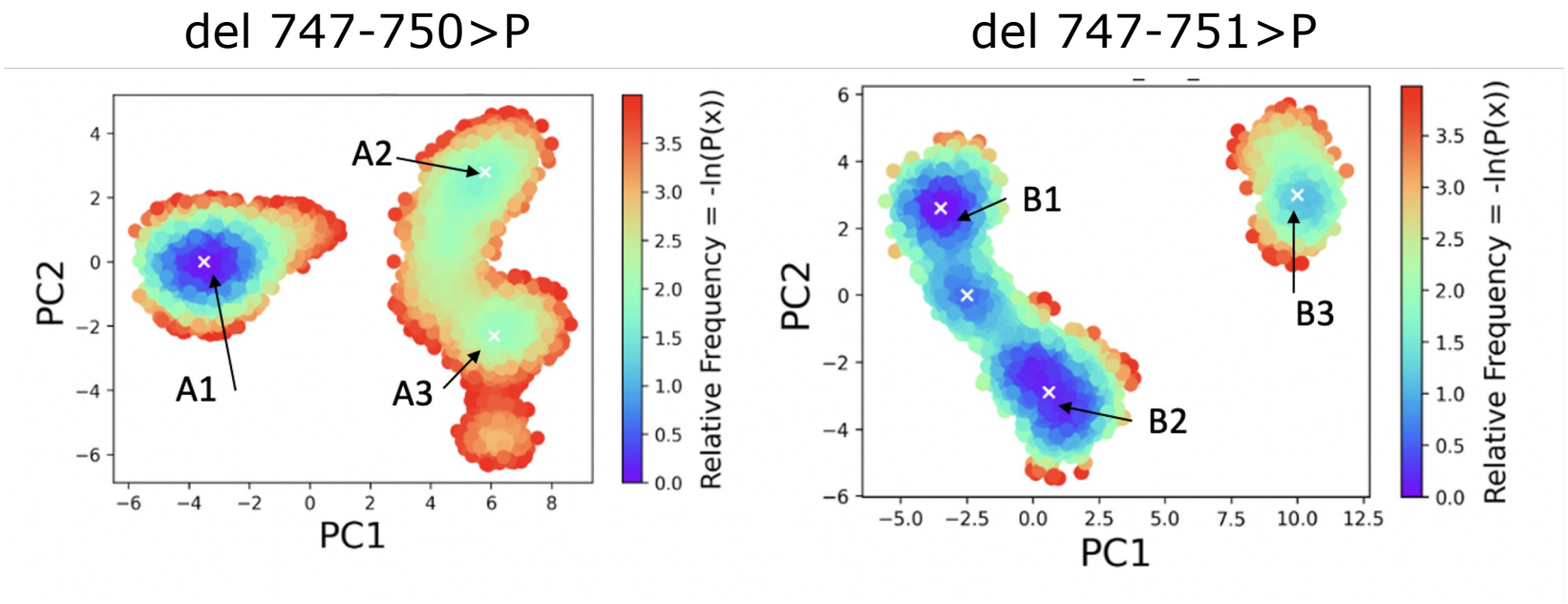

Fig 1A. Projections of the conformations visited in the MD trajectories on the top two principal components with relative frequencies colored according to the color bar for a profile 1 variant: del 747-750>P and a profile 2 variant: del 747-751>P differing by just one amino acid.

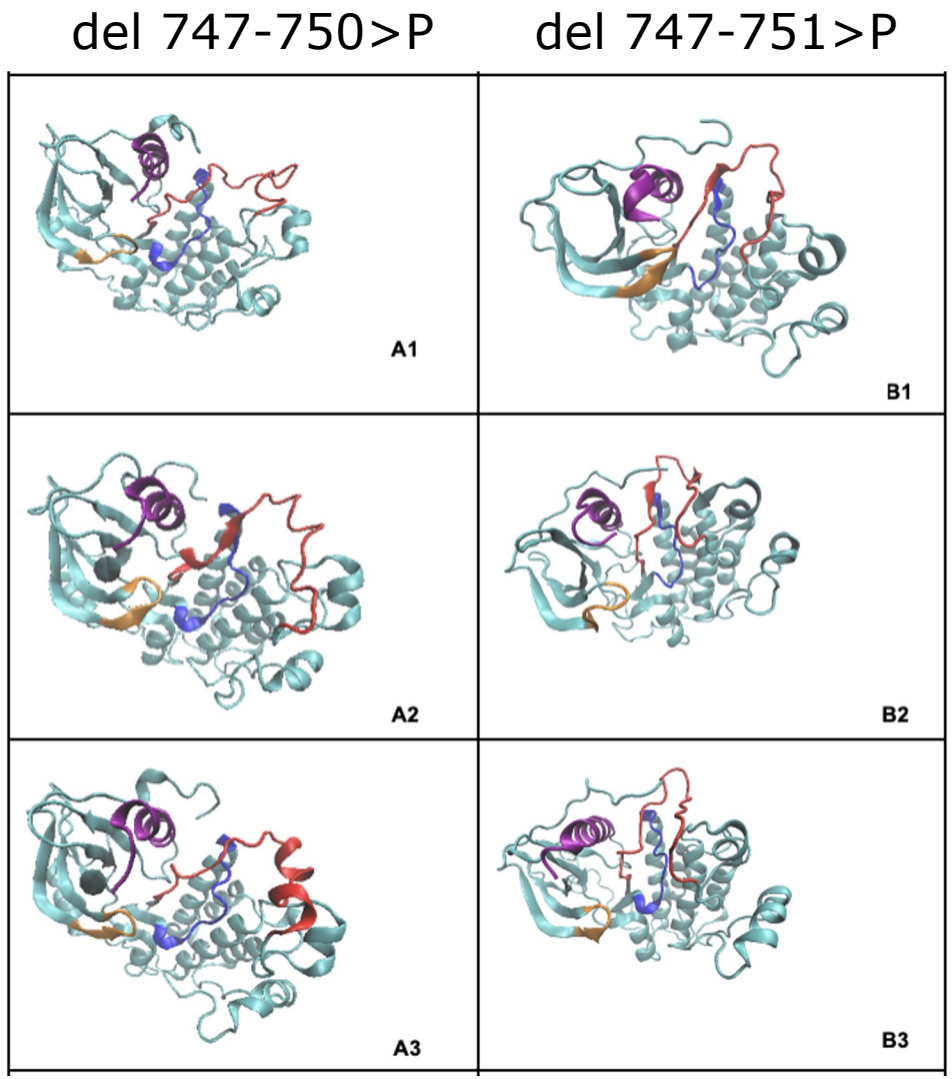

Fig 1B. Structures extracted from the PCA clusters (A1-A3 and B1-B3) as marked in Fig 1A. The profile 1 variant predominantly shows localized motion of the activation loop and the alpha C helix. Whereas the profile 2 motion shows delocalized motion involving other subdomains in the C-lobe and the N-lobe.

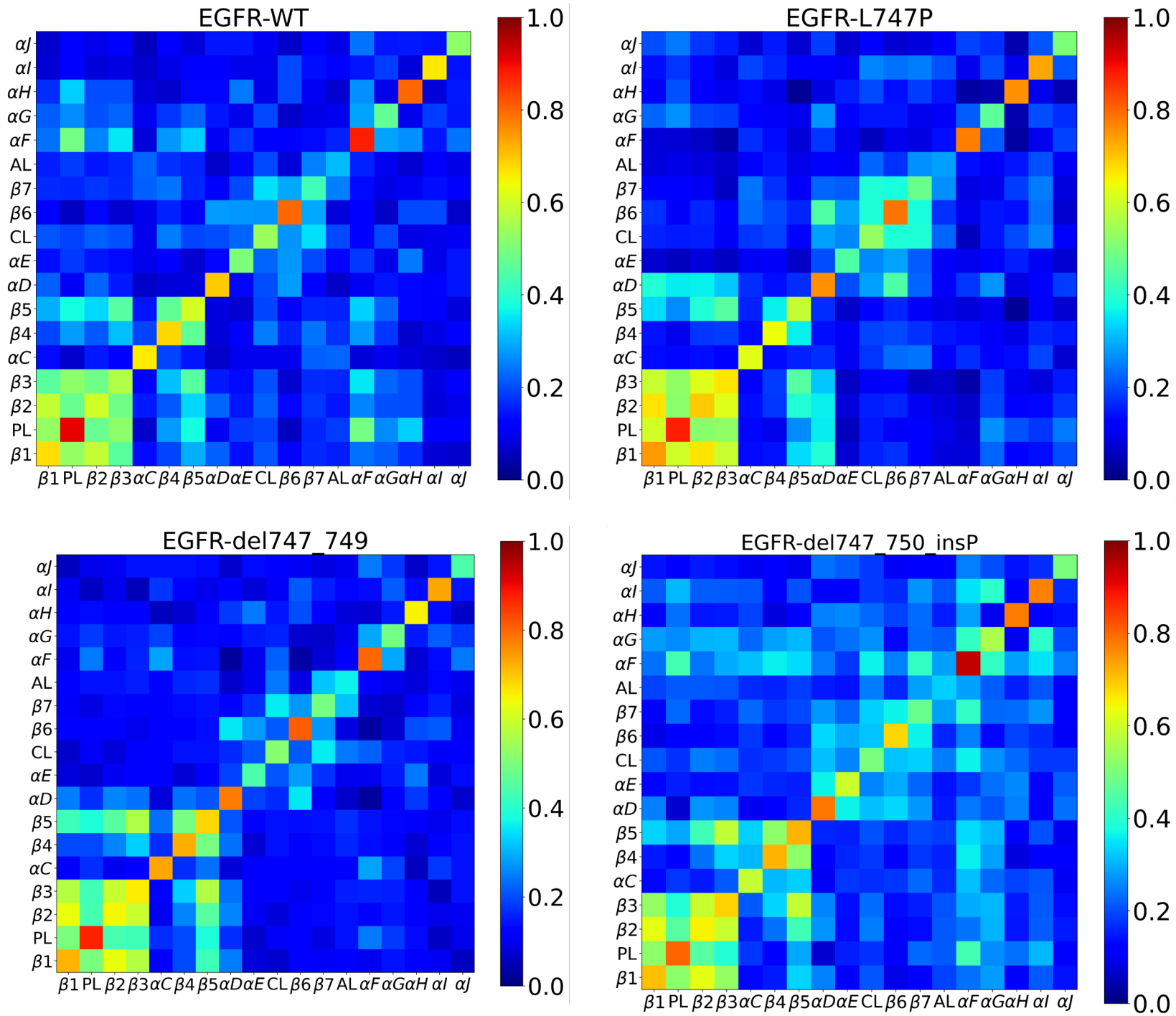

Fig 2A. Coarse-grained dynamical cross-correlation matrices of EGFR Exon-19 deletion mutants from profile 1.

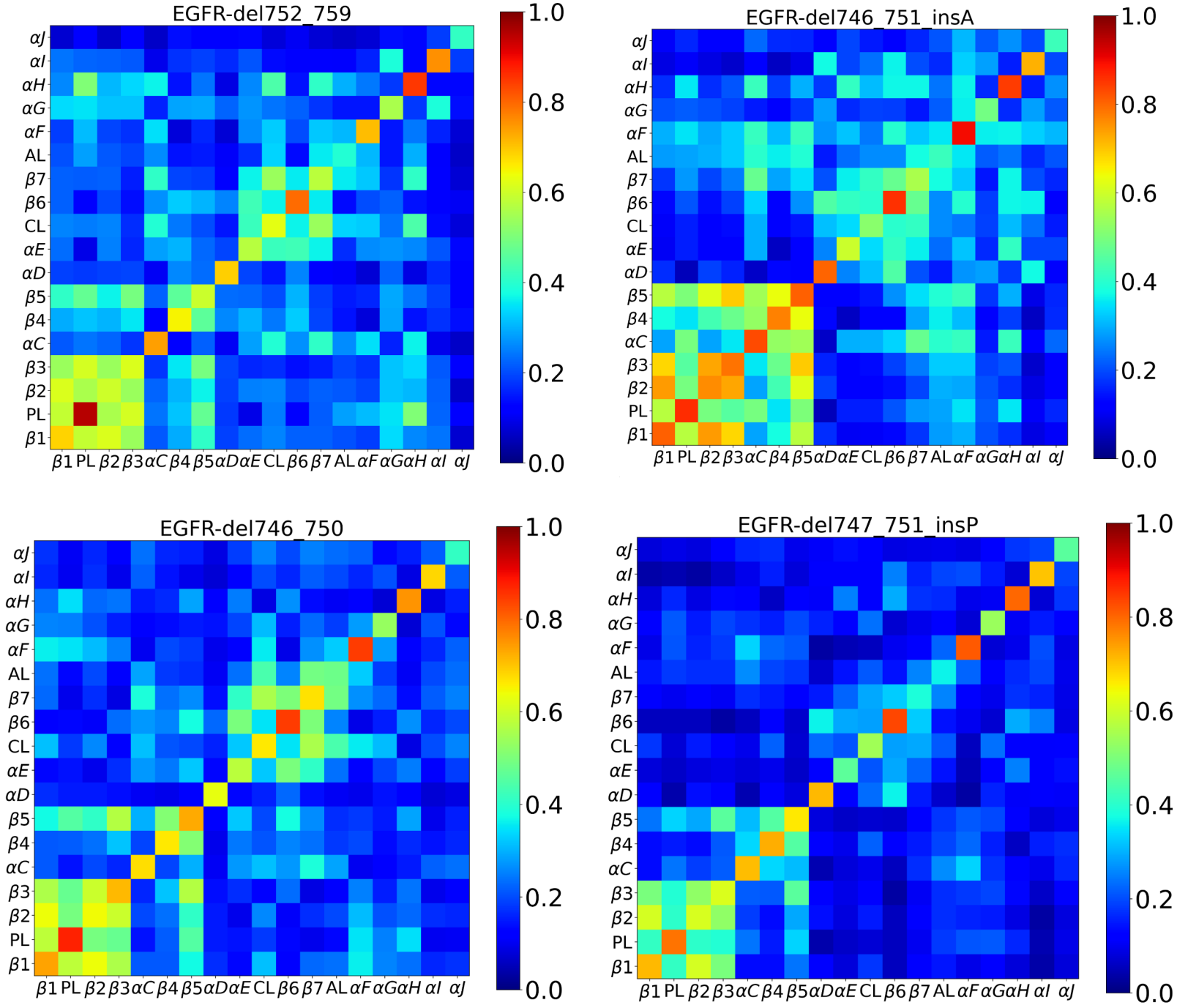

Fig 2B. Coarse-grained dynamical cross-correlation matrices of EGFR Exon-19 deletion mutants from profile 2.

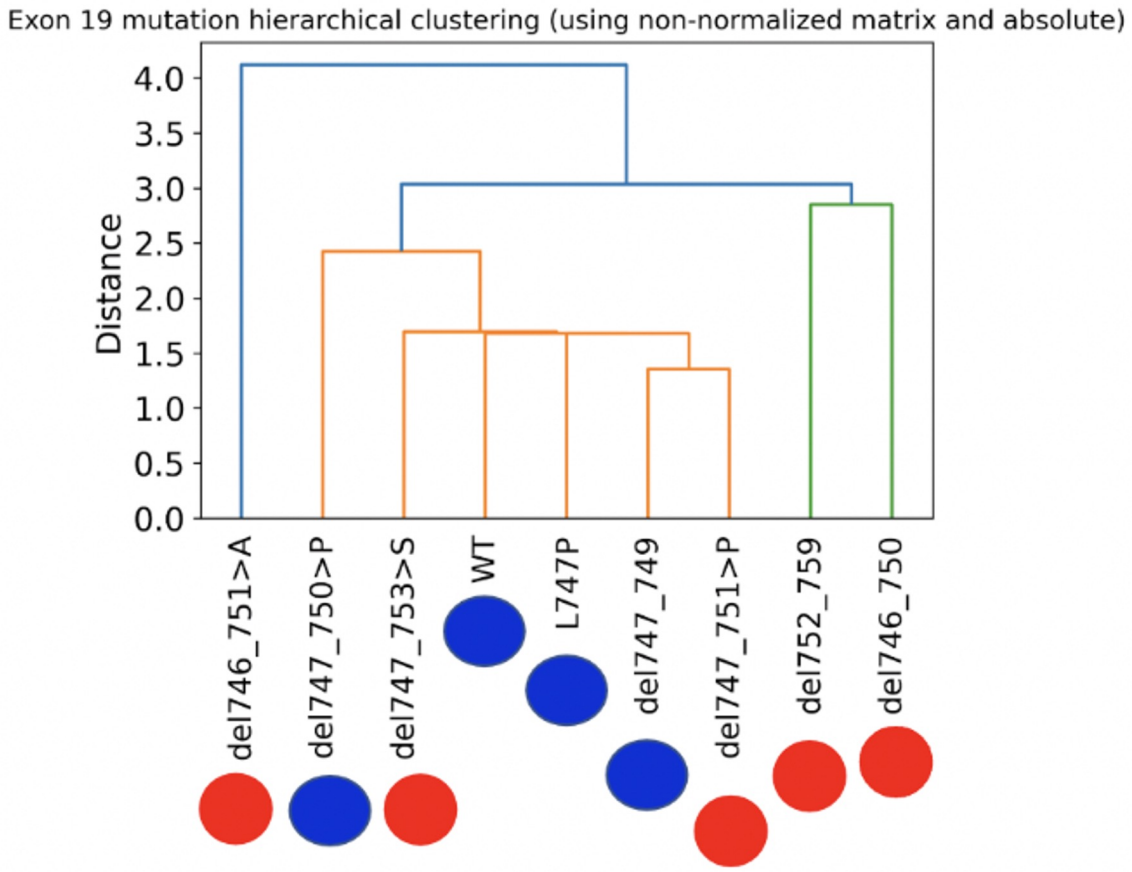

Fig 2C. Hierarchical clustering of the EGFR Exon-19 deletion mutants. Mutations from profile 1 are marked in blue and mutations from profile 2 are marked in red.

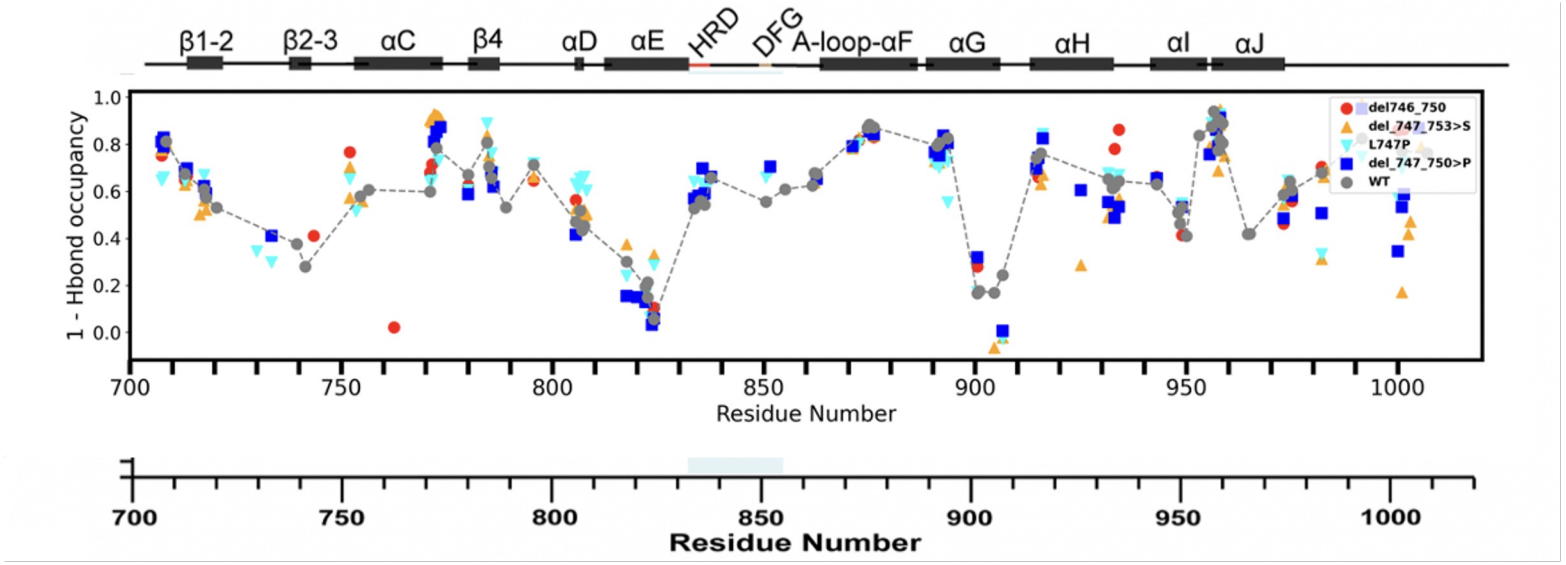

Fig 3A. Hbond occupancy per residue obtained from simulations.

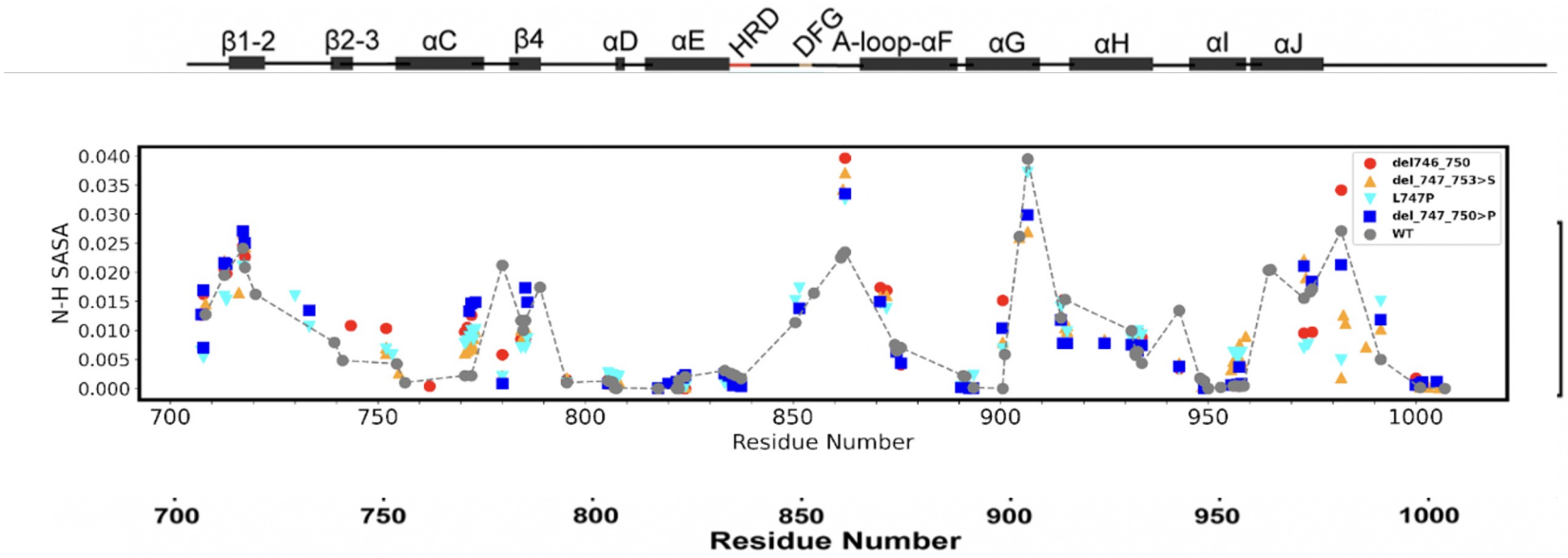

Fig 3B. Solvent accessible surface area or SASA per residue obtained from simulations.

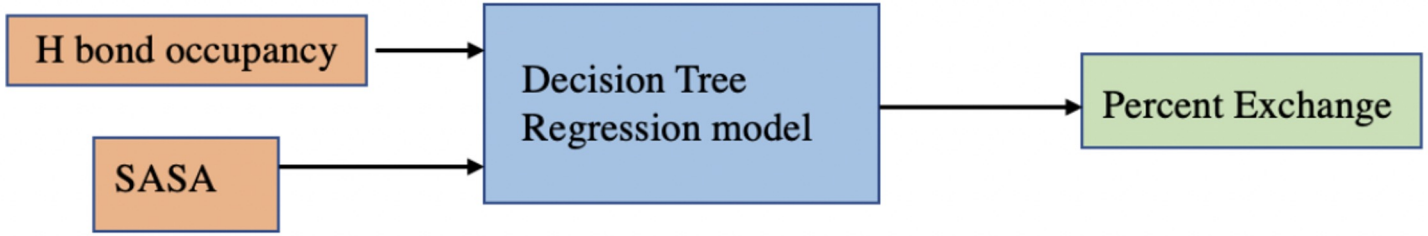

Fig 3C. Decision Tree regression model trained with Hbond occupancy and SASA as inputs to predict Percent Exchange

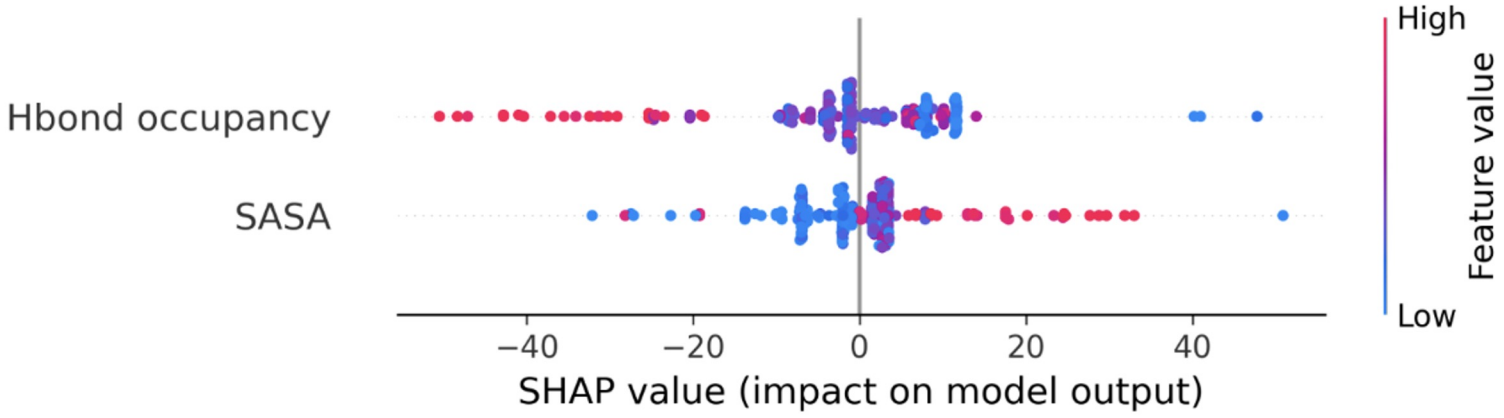

Fig 3D. SHAP values determine the contribution from the Hbond and SASA towards % exchange, while also quantifying the nature of contribution.

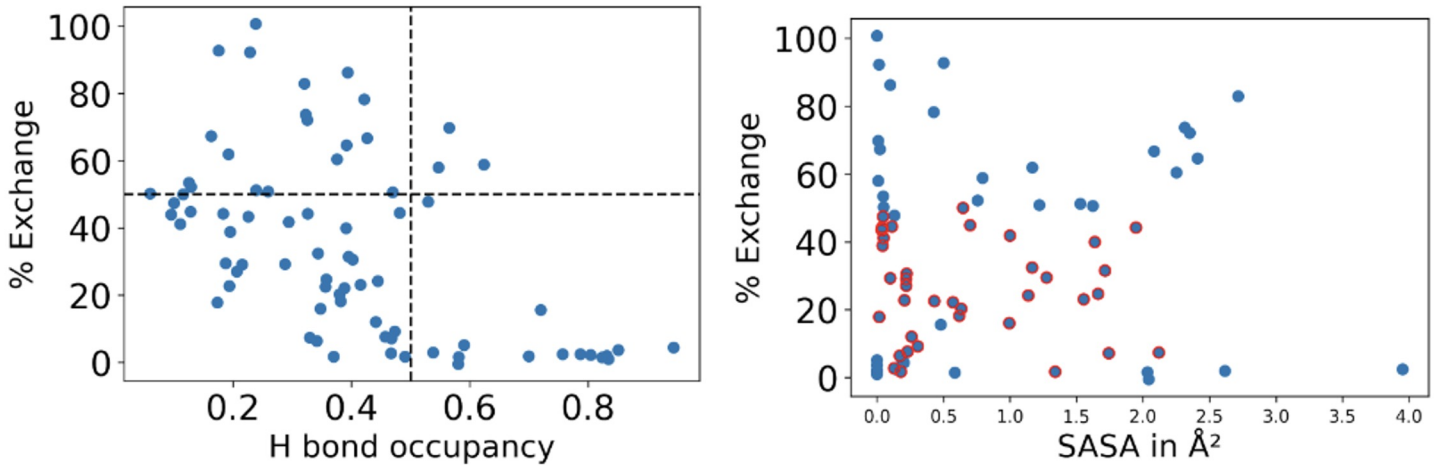

Fig 4A. % exchange versus Hbond occupancy and versus SASA for the peptides measured in HDX for wildtype EGFR.

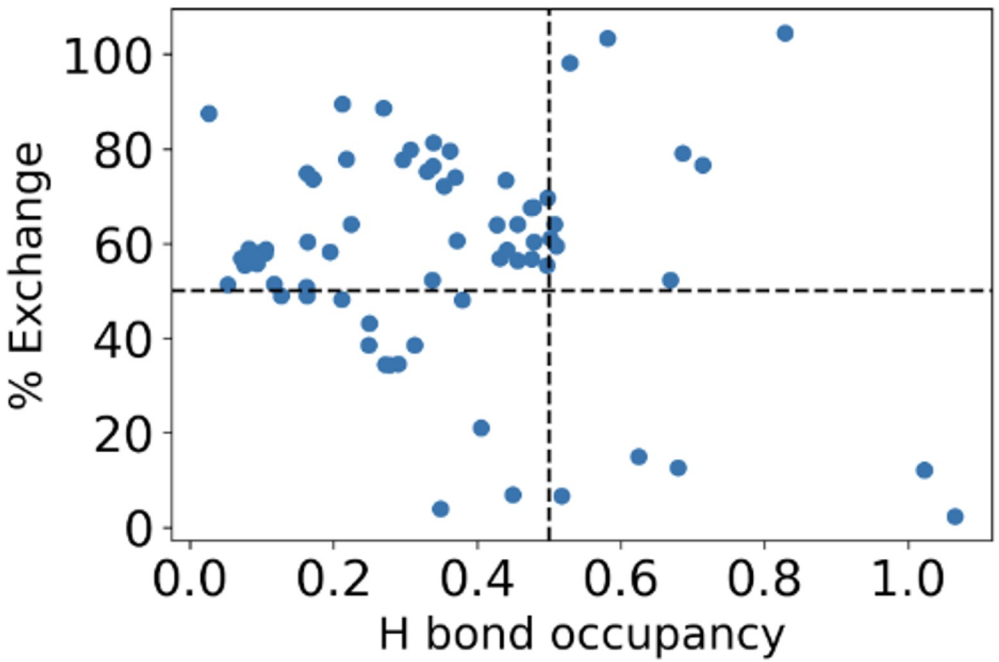

Fig 4B. % exchange versus Hbond occupancy and versus SASA for the peptides measured in HDX for del747-753-insS EGFR.

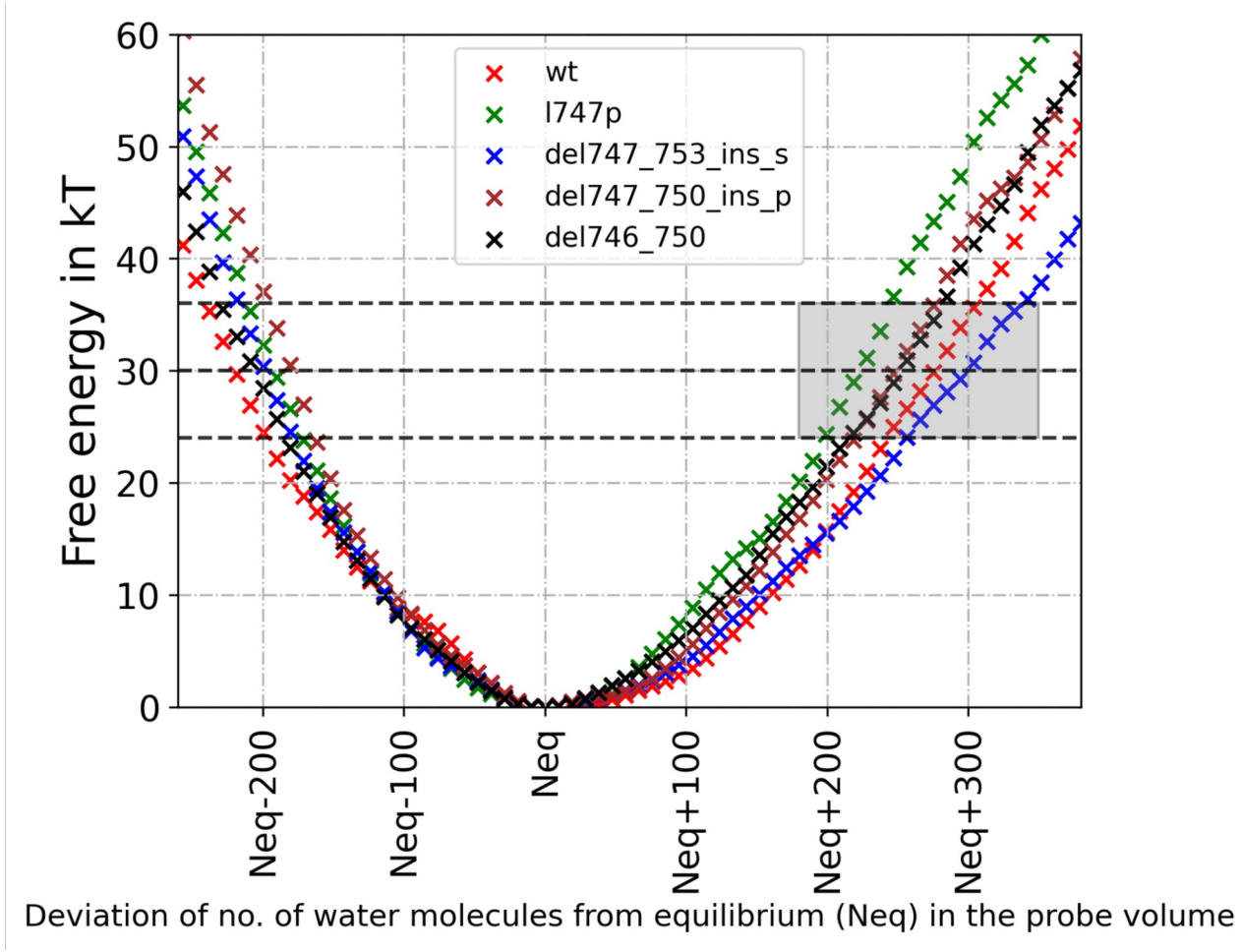

Fig. 4C. Free energy profiles obtained through INDUS sampling.

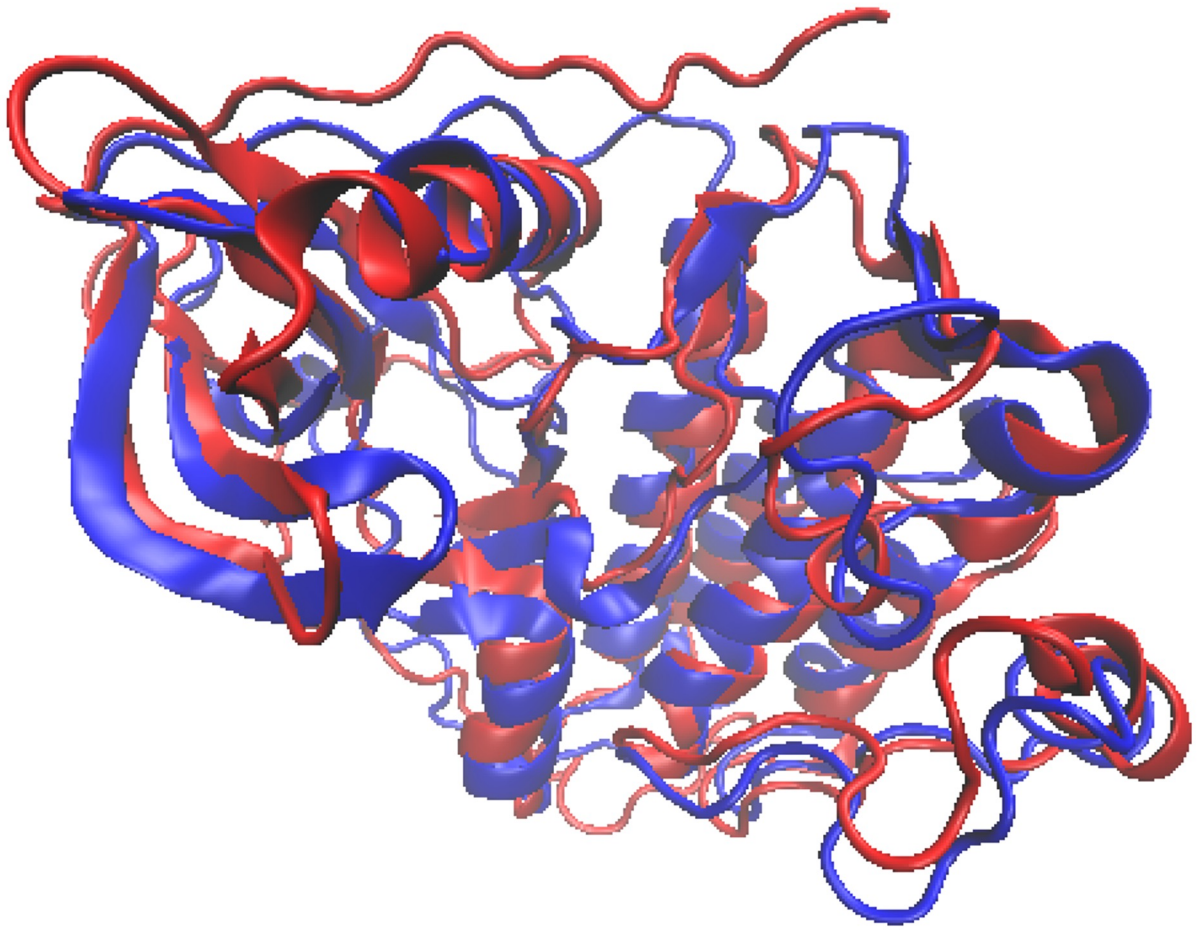

Fig 4D. Structural alignment between average structure for del-747-753-insS obtained from unbiased MD (shown in red) and INDUS corresponding to the free energy barrier of 30 k_BT (shown in blue).

A

EGFR	Ligand	Combind Score	Glide Gscore	delta (ATP-Erlotinib)
WT	ATP	3.515	-6.407	-1.39
	Erlotinib	3.437	-5.019	
L747P	ATP	1.510	-5.096	-1.605
	Erlotinib	2.453	-3.491	
del747_750_insP	ATP	3.605	-7.554	-2.581
	Erlotinib	3.718	-4.973	
del746_750	ATP	3.153	-5.544	1.153
	Erlotinib	6.579	-6.697	
del747_753_insS	ATP	2.804	-5.781	0.414
	Erlotinib	6.797	-6.195	

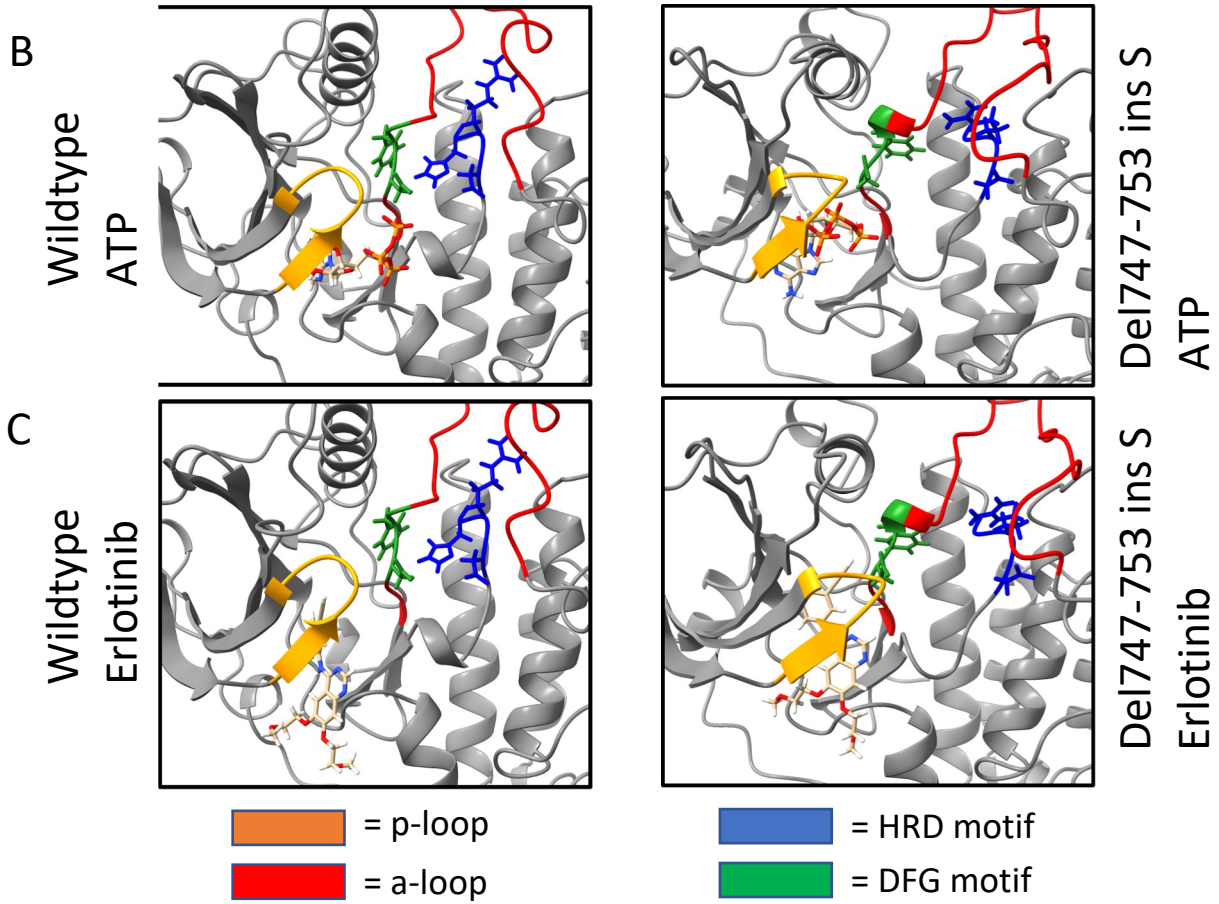

Fig 5. A. Results from molecular docking of ATP and erlotinib for EGFR variants. The zone that yielded the best docking score, which also conserves the known binding mode of the ligands is reported. B, C. Snapshots of docked ATP and erlotinib for representative profile 1 and profile 2 mutants are depicted.

		P loop	Alpha C helix	Activation Loop	Catalytic Loop	N terminal	C terminal
WT	PC1	M	M	H	0	0	0
	PC2	H	H	H	0	0	0
L747P	PC1	M	0 (always "in")	M	0	0	0
	PC2	M	0 (always "in)	M	0	0	0
del 747-750-insP	PC1	M	0 (always "in")	M	M	0	0
	PC2	M	0 (always "in")	0	0	0	0
del 747-749	PC1	M	H	0	0	M	M
	PC2	M	H	M	M	M	M
del 746-750	PC1	M	0(always "in")	0	0	0	H
	PC2	M	M	0	0	H	0
del 747-753-insS	PC1	0	0(always "in")	0	0	H	0
	PC2	M	0(always "in")	0	0	M	0
del 746-751-insA	PC1	M	0(always "in")	H	0	0	0
	PC2	M	M	H	M	0	H
del 752-759	PC1	H	0(always "in")	M	M	0	H
	PC2	0	0(always "in")	M	0	0	H
del 747-751-insP	PC1	M	M	H	H	M	M
	PC2	M	M	M	0	0	H

Profile 1

Profile 2

Table 1: Observed motion projected along the top two principal components in several subdomains for EGFR deletion mutants classified as high (H), medium (M), and low (0).

β1 – 712 to 720	Catalytic loop – 835 to 842
P loop – 719 to 724	β6 – 843-847
β2 – 724 to 731	β7 – 849 to 853
β3 – 740 to 747	Activation loop – 855 to 876
αC – 756 to 767	αF – 883 to 888
β4 – 777 to 782	αG – 893 to 909
β5 – 786 to 791	αH – 923 to 929
αD – 798 to 806	αI – 941 to 950
αE – 811 to 831	αJ – 973 to 985

Table 2: Subdomains and key sub-regions of the EGFR kinase domain. These subdomain definitions are used to compute the dynamical cross correlation matrices.

WT and Exon-19 Deletion Mutant	Mutual Information	Joint Mutual Information	Normalized Mutual Information
WT	1.4	8.14	0.42
L747P	1.11	8.08	0.34
del747_750_insP	1.20	8.13	0.36
del746_750	1.62	7.81	0.49
del747_753_insS	1.34	7.5	0.4

Table 3: Mutual information values of EGFR deletion mutants computed from molecular dynamics simulations.

Protein	Spring constant, k (kJ Mol^{-1})		Biasing simulation length (ns) per window
WT	0.012		10
L747P	0.015		10
Δ747-750>P	0.010		10
Δ747-753>S	0.0111		10
Δ746-750	0.015		10

Table 4: Parameters related to the INDUS enhanced sampling simulations for EGFR variant systems.