High-Throughput Prediction of the Impact of Genetic Variability on Drug Sensitivity and Resistance Patterns for Clinically Relevant EGFR Mutations from Atomistic Simulations

Aristarc Suriñach[1#], Adam Hospital[2#], Yvonne Westermaier[1#], Luis Jordà[3], Sergi Orozco-Ruiz[3], Daniel Beltrán[2], Francesco Colizzi[2%], Pau Andrio[3], Robert Soliva[1], Martí Municoy[1], Josep Lluís Gelpí[3,4], and Modesto Orozco[2,4]*

[1] Nostrum Biodiscovery, Av. Josep Tarradellas 8-10, 08029 Barcelona, Spain
[2] Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona 08028, Spain.
[3] Barcelona Supercomputing Center (BSC)
[4] Department Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain.
* Corresponding author: modesto.orozco@irbbarcelona.org
[#] Equally contributing authors.
[%] Current address: Institute of Marine Sciences, ICM-CSIC, Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain

ABSTRACT

Mutations in the kinase domain of the Epidermal Growth Factor Receptor (EGFR) can be drivers of cancer and also trigger drug resistance in patients receiving chemotherapy treatment based on kinase inhibitors. *A priori* knowledge of the impact of EGFR variants on drug sensitivity would help to optimize chemotherapy and design new drugs that are effective against resistant variants before they emerge in clinical trials. To this end, we explored a variety of *in silico* methods, from sequence-based to 'state-of-the-art' atomistic simulations. We did not find any sequence signal that can provide clues on when a drug-related mutation appears, or the impact of such mutations on drug activity. Low-level simulation methods provide limited qualitative information on regions where mutations are likely to cause alterations in drug activity, and they can predict around 70% of the impact of mutations on drug efficiency. High-level simulations based on non-equilibrium alchemical free energy calculations show predictive power. The integration of these 'state-of-the-art' methods into a workflow implementing an interface for parallel distribution of the calculations allows its automatic and high-throughput use, even for researchers with moderate experience in molecular simulations.

INTRODUCTION

EGFR (Epidermal Growth Factor Receptor) overexpression, or certain mutations in both ecto-domain or Tyrosine Kinase domain can lead to impaired activation of its tyrosine kinase (TK) activity, thereby triggering the hallmarks of cancer, i.e., increased proliferation and survival of tumour cells, aggressive invasion and metastasis, evasion of cell death, and increased metabolism [1, 2]. Unregulated EGFR over-activity is present in many types of cancer, especially in those with the poorest prognosis. For example, over 60% of patients with metastatic non-small-cell lung cancer (NSCLC) overexpress EGFR [3, 4], and around 10-30% [5-7] of these individuals have mutations leading to ligand-independent activation of EGFR TK. Some of these activating mutations are located in the extracellular domain and drive a conformational change from the inactive to the active state, mimicking that induced by the natural ligand (the epidermal growth factor (EGF) [8]). Others are located at the kinase domain and trigger changes in the ATP-binding pocket that alter the 'on-off' equilibrium of the enzyme [9-13], probably by destabilizing the 'inactive state' with respect to the 'active one' [14, 15].
Current treatments for cancers involving EGFR dysregulation are based on either monoclonal antibodies directed against the cognate ligand EGF or EGFR's homo- or heterodimerization, or on inhibitors of TK activity. Various structurally related FDA-approved small molecules either reversibly or irreversibly compete with the natural substrate ATP to inhibit TK activity. Several of these drugs are used to treat non-small cell lung, pancreatic, colorectal, head and neck, and breast cancers [11, 16, 17]. The first TK inhibitors (TKIs) with clinical benefits were **Erlotinib** (Tarceva®) and **Gefitinib** (Iressa®) [18-21], and third-generation **Osimertinib** [22, 23], which forms a covalent bond with C797 after an initial non-covalent binding. Unfortunately, while these drugs show good antineoplastic activity at the beginning of the treatment, drug resistance appears as cancer progresses and tumour cells accumulate mutations. This resistance is triggered by the emergence of inactivating mutations [24], which rescue the 'dysregulated' activity of EGFR [25-30]. Several drug-inactivating mechanisms have been proposed, including, among others, the activation of alternate proteins downstream of EGFR signalling, the activation of proteins that feed into the EGFR signalling cascade, and a decrease in the affinity of TKIs [31].

While resistance driven by the rewiring of cellular networks is independent of the molecular details of the EGFR inhibitor and can be tackled by multidrug therapies [32], resistance caused by mutations with decreased inhibitor affinity are dependent on the fine details of the drug and the mutation and are susceptible to theoretical predictions by means of simulation methods. Here we focus on mutations in the kinase domain that can affect inhibitor binding in a drug-specific manner. In fact, first-generation ATP-competitive inhibitors quickly faced TK mutation-induced treatment resistance related to a decrease in inhibitory potency, which drove the development of second- and third-generation inhibitors [33]. Unfortunately, while improving resiliency to mutations, second-generation inhibitors show limited efficacy in circumventing some mutations, and even third-generation inhibitors are susceptible to inactivating mutations that affect the vicinity of the ATP binding site [34].
Improvements in EGFR-TKI-based therapies would require detailed knowledge of the impact of mutations on the activity of the drugs. Patient genotyping [35], followed by *in silico* predictions and *in vitro* validation, could help oncologists ascertain whether mutations render the kinase domain (KD) of EGFR resistant to therapeutic drugs. Furthermore, *in silico* prediction of the impact of mutations on kinase inhibition would not only allow an understanding of the impact of known mutations on existing drugs but would also help to predict resistance and implement modifications in the therapy before relapse happens. Even more exciting is that *in silico* mutagenesis and binding predictions would allow pharmaceutical laboratories to anticipate inactivating mutations for a drug candidate before it reaches the market, thereby helping to stratify patient cohorts in clinical trials and triggering the development of a modified drug candidate able to evade inactivating

mutations. To this end, a reliable simulation-based pipeline has to be developed and implemented in a user-friendly manner for non-experts.

Here we present a multilevel and automated approach that allows the *in silico* prediction of the effect of mutations on the binding properties of TKIs targeting EGFR. A variety of sequence-based methods helped to detect some trends of drug-affecting mutations but are far from having any predictive power. Simulation methods explicitly accounting for the structural and dynamic properties of the protein and the drug-protein complex achieve predictive power ranging from 70% for the lower-level methods to an impressive 100% for the most elaborate methods based on molecular dynamics and non-equilibrium free energy calculations. The implementation of these state-of-the-art techniques in an automated workflow involving highly parallel computers allows non-expert users to perform these calculations in hours with moderate computational resources, minutes with a pre-ExaScale parallel supercomputer, and seconds with an ExaScale one.

METHODS

Dataset: Considerable data on somatic cancer mutations have been gathered via sequencing projects. For this study, sequence variants found in samples from cancer patients were extracted from the International Cancer Genome Consortium [36] (ICGC; https://dcc.icgc.org/), the Catalog Of Somatic Mutations In Cancer [37] (COSMIC; https://cancer.sanger.ac.uk/cosmic), and the Clinical Variants [38] (CLINVAR; https://www.ncbi.nlm.nih.gov/clinvar/) databases. For EGFR, the ICGC database yielded 5710 somatic mutations, of which 5158 were substitutions, including 402 missense substitutions. Only variants leading to changes in the anticancer activity of TKIs were retained (Table 1). We searched the literature for the annotated origin of resistance (Suppl. Table S2), and mapped the key regions for binding and activity onto the structure of the TK domain, which comprises the nucleotide-binding loop (P-loop), the catalytic loop (C-loop), the αC-helix at the dimerization interface, the activation loop (A-loop), the hinge, and the DFG triad (see Figure 1).

Sequence analysis: We used sequence alignments to determine whether mutations affecting drug-binding are located in variable or conserved regions and whether such mutations are common even in the absence of the evolutionary pressure of the drug. To this end, we extracted the sequences of 94 human TKs from KinBase [39], aligning them with ClustalW as implemented in the msa R package [40]. ClustalW [41, 42] is one of the widest used programs for multiple alignment, it works generating pair-alignments from which phylogenetic trees are created and used as reference for the multiple alignment. Sequence variability at each position of the TK domain was determined from the Shannon entropy score, as described elsewhere [43] (see Eq. 1):

$$H = -\sum_{i=1}^{M} P_i log_2 P_i \qquad (1)$$

where the sum extends for all mutations sampled in the alignments at a given position, and P_i stands for the probability of residue i at a given position. Non-pathological human polymorphisms mapping onto the TK domain of EGFR were extracted from the gnomAD database [44].

In order to evaluate the likelihood of an aminoacid substitution (wild type → mutant) we used 20x20 BLOSUM62 matrices [45]. A high positive value of a BLOSUM62 index associated to a mutation X → Y means that these mutations are commonly found in proteins (in multiple alignments), while large negative values mean that these are rare changes in proteins (expected disruptive mutations).

Pathogenicity analysis: We used the PMUT program [46, 47], ranging from sequence and structural information to the protein [48, 49], to evaluate the pathological potential of mutations affecting drug response. PMUT uses a large number of sequence-dependent parameters (e.g. local and global conservation, predicted structural elements, aggressiveness of the mutation) and a Machine Learning algorithm (Random Forest) trained to

distinguish between neutral and mutations associated to Mendelian diseases. It is one of the most used programs for pathogenicity prediction and is freely available at https://mmb.irbbarcelona.org/PMut/

Modelling of the complexes: We used the structure of PDB entry 4WKQ as a template, using PDB entry 2ITY to fill the structural gaps. Both structures contain **Gefitinib** (Iressa®), covering residues 694 to 1020 at 1.85 Å resolution (4WKQ), and residues 697 to 1019 at 3.42 Å resolution (2ITY). Refinement involved checking alternative conformations per residue in the 4WKQ parts, keeping only those with higher occupancy. For the crucial D855 in the DFG-motif, we chose as a starting point the side-chain orientation that forms a double salt bridge with the catalytic K745 and E762 in the αC-helix. The structure of the final model was checked using our local *BioBB* validation module [50]. The binding geometry of **Erlotinib** was taken from 4HJO, **Lapatinib** from 1XKK, and **Osimertinib** from 4ZAU. The binding geometry of **Icotinib** was taken from the *BioBB* AutoDock-Vina [51] module, defining the binding pocket as those residues closest to 6.5 Å from **Gefitinib** in the 4WKQ structure. In all cases, structural water molecules present in the crystal around the binding site were maintained, the hydrogen atoms being oriented during the molecular dynamics (MD) setup procedure as discussed below. Binding interactions of the different drug molecules with the modelled structure of this study after minimization (relaxation of bad contacts) are shown in Suppl. Figures S1 to S5. A comparison between the modelled structures of this study after minimization (relaxation of bad contacts) and experimental structures of the PDB bearing the same mutated amino acids and ligands is also included in Suppl. Figure S6.

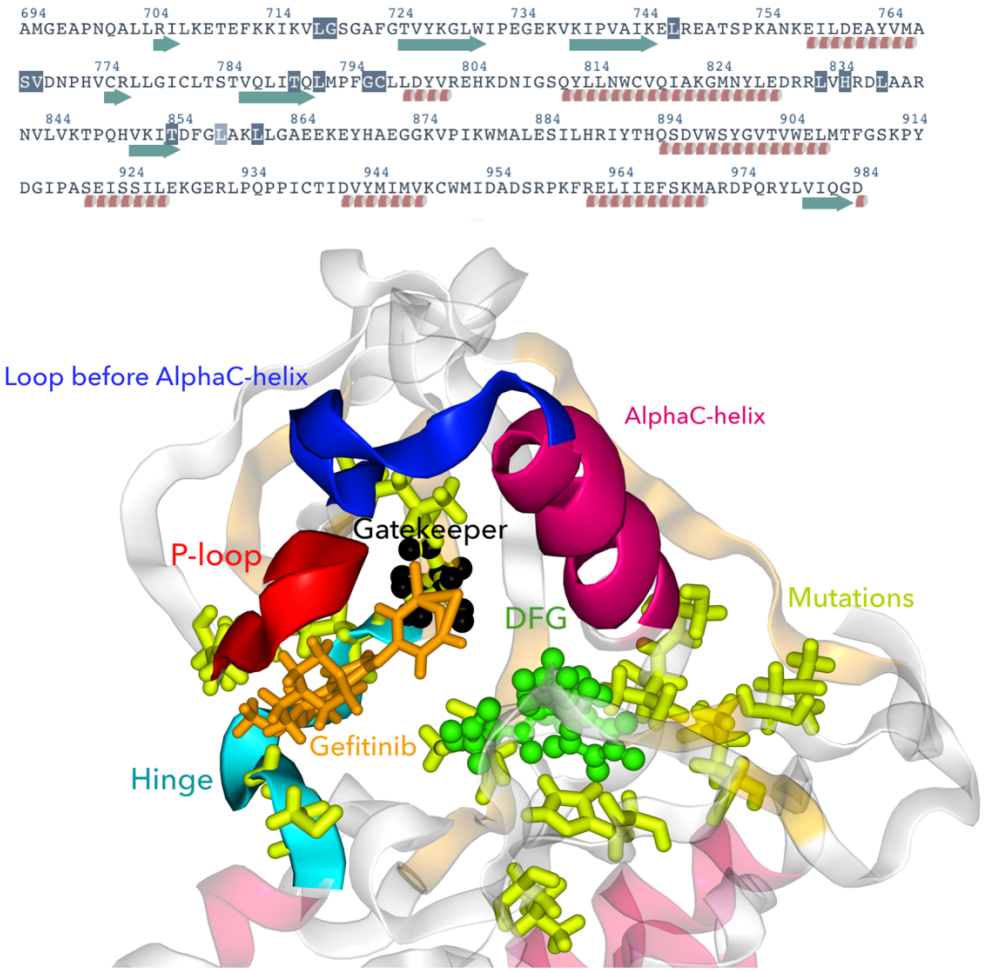


Figure 1: *Location of clinically relevant mutations in the kinase domain of EGFR. Top: TK domain sequence with the positions of the studied mutations highlighted. Bottom: Representation of the active centre with the studied mutations and the most important regions on the protein structure. Clinically Relevant Mutations: Yellow Licorice; Gefitinib (IRE): Orange Licorice; Gatekeeper: Black CPK; Hinge: Light Blue Cartoon; P-loop: Red Cartoon; DFG motif: Green CPK; Loop before AlphaC-helix: Dark Blue Cartoon; AlphaC-helix: Violet Cartoon; PDB Structure: 4WKQ as a template, 2ITY used to fill the structural gaps. The 3D representation can be interactively explored in the 3dRS server: https://mmb.irbbarcelona.org/3dRS/s/MZplaW*

Force-Field parameters: Proteins were described by the AMBER99SB-ILDN force field [52], water by the TIP3P model [53], and counterions by the AMBER99SB-ILDN-associated ion model. For ligands, we used our web-based automatic tool, taking care to define the suitable charge state of the ligands (suitable charge state and parameters were obtained using an automated protocol [54-56], which relies on charges determined from fitting to electrostatic potential, van der Waals transferred from AMBER99SB-ILDN types and GAFF [57], and torsions refined by an automated protocol using DFT/SCRF calculations as a reference [54-56]).

Generation of initial mutant geometries: Starting from the wild-type (WT) geometries above, the *biobb_structure_checking* module was used to create and validate the geometry of the mutants. The structure-checking steps covered proper amide assignments, chirality, cis/trans backbone, disulphide bridges, and severe intra-protein clashes. Neither the WT nor any of the modelled mutants presented any major issue. In all cases, models were subjected to minimization, thermalization, and 50 ns equilibration before production. For each mutant, the residue protonation state was defined at a physiological pH of 7.4 using the *PROPKA* software (v3.1.8) [58], with reorientation of the side chains of histidine residues using the PDB2PQR (v2.1.1) webserver (https://server.poissonboltzmann.org/pdb2pqr) [59].

Induced fit calculations: We used the PELE suite of programs [60] to analyse the docking of ligands to WT and mutant proteins in an unbiased manner. PELE uses a Metropolis-Monte Carlo/annealing protocol that combines ligand random moves, main chain perturbation based on normal mode displacements, and a final relaxation stage consisting of side-chain relaxation and global minimization. PELE uses a continuum, but accurate solvent, and a broad definition of the binding site. Compared with other docking algorithms it is slower, but allows a very exhaustive exploration of the binding landscape, being one of the most accurate docking-like methods in cases of induced fit. We compared the docking energies of the WT and mutant to detect mutation-induced changes in binding. Crystal structures (see above) for the WT and mutants were built using Schrodinger's Protein Preparation Wizard [61] and Maestro [62], defining a docking sphere of a radius of 5 Å around the ATP binding site. Defaults were used for PELE. OPLS2005 [63] was used to define the energy functional combined with the SGB solvent model corrected for non-polar contacts [64]. Rotamer libraries were taken from our Peleffy library. Binding poses were obtained using AdaptivePELE [65] to explore diverse binding poses within the docking sphere. The resulting poses were then clustered, and the most relevant ones were re-explored. A typical PELE calculation takes around 24 hours in a 64 core Intel-based computer.

Molecular Dynamics (MD) simulations: For each apo and holo EGFR variant, ten replicas were generated. Each variant was optimized to relax bad contacts and solvated in dodecahedral water boxes extending for at least 12 Å from any atom of the protein. Counterions (Na^+ and Cl^-) were added to maintain neutrality including additional ions to adjust to a 150 mM salt concentration. Solvated systems were then reoptimized, and slowly thermalized (310 K) and equilibrated, first in the NVT ensemble for 1 ns before moving to the NPT one, slowly removing restraints on the protein and the ligand heavy atoms for 1 ns. Each replica of the final systems was relaxed for 50 ns in the NPT ensemble (P=1 atm; T=310 K) before production runs (100 ns each replica), from which seeds for forward and reverse free energy calculations were extracted (see below). Newton's equations of motion were integrated every 2 fs using LINCS [66] to maintain all bonds involving hydrogen frozen at equilibrium distances. Periodic boundary conditions and Particle Mesh Ewald methods [67, 68] were implemented to capture long-range interactions. Parrinello-Rahman thermostats/barostats [69, 70] were used to maintain the pressure and temperature at desired values. All MD calculations were done using GROMACS (v2018.4) [71].

Interaction profiles: We used the collected trajectories of the complexes to obtain the residue-drug interaction energies and those residues with the strongest interactions [72], as determined from the combination of electrostatic (computed using Poisson-Boltzman calculations [73]) and van der Waals interactions. Group fragmentation was done to maintain monopoles, as describe elsewhere [73].

MMPBSA calculations: Molecular Mechanics with Solvent-Accessible Surface Area correction [74] was performed in pilot calculations, considering the bound and unbound state of the WT and mutant proteins and the standard defaults, as described elsewhere [75].

Machine Learning (ML) predictors. We explored the use of PremPLI [76], a method that uses a series of structural descriptors derived from the complex geometry and a random forest classifier to predict when a mutation in a protein impacts the binding of a drug. The method has been recently shown [76] to outperform other ML-based methods (mCSM-lig [77], Aldeghi ML1 [78]), and can be accessed via a public and well maintained website (https://lilab.jysw.suda.edu.cn/research/PremPLI/).

Free energy calculations: The mutation-induced change in the binding free energy for the different inhibitors was computed using standard thermodynamic cycles, comparing the free energy change associated with the mutation in the apo and drug-bound protein states (Figure 2). Individual free energies were computed via non-equilibrium methods: the *Crooks Gaussian Intersection (CGI)*, *Jarzinsky's equality (JE)* and the *Bennett Acceptance Ratio (BAR)* methods [80]. Contrary to free energy perturbation or thermodynamic integration (TI), non-equilibrium methods determine the free energy of an alchemical process from the distributions of irreversible work caused by a system change (an amino acid mutation in our case), obtained in 'forward' and 'reverse' directions (see Figure 3). We generated a meta-trajectory concatenating the 10 collected replicas for the WT and mutant in both the apo and holo states (see Figure 2), selecting 100 random configurations as starting points for slow-growth TI non-equilibrium perturbations using de Groot's PMX protocol [79-81]. For each mutation, we ran 100 (replica) x2 (wild type → mutant and mutant → wild type) x2 (bound and unbound states) alchemical changes. Each of the 100 perturbation TI trajectories were extended for 50 ps after a series of test analyses had demonstrated it to be a good compromise between accuracy and computational efficiency (see Suppl. Figure S7 and reference [82]). The histograms of irreversible work for the forward and reverse transitions were calculated by the three methods outlined before (see Figure 2) to determine three independent estimates of the reversible free energy associated with the mutation. When a discrepancy between the three estimates was large (standard deviation of more than 3 kJ/mol, with one of the estimates leading to a global change in the predicted free energy change), simulations were extended to 500x2x2 alchemical changes to check for convergence between the different estimates. With this simulation set-up the major origin of uncertainty arises from no purely Gaussian nature of the irreversible work distribution, which challenges the accuracy of the CGI method. Typical CPU times for MD/PMX workflow implies around 2x130 hours in a 64 cores computer, and thanks to the extreme parallelism provided by the workflows only 2x8 hours in a medium sized 1024 cores computer cluster.

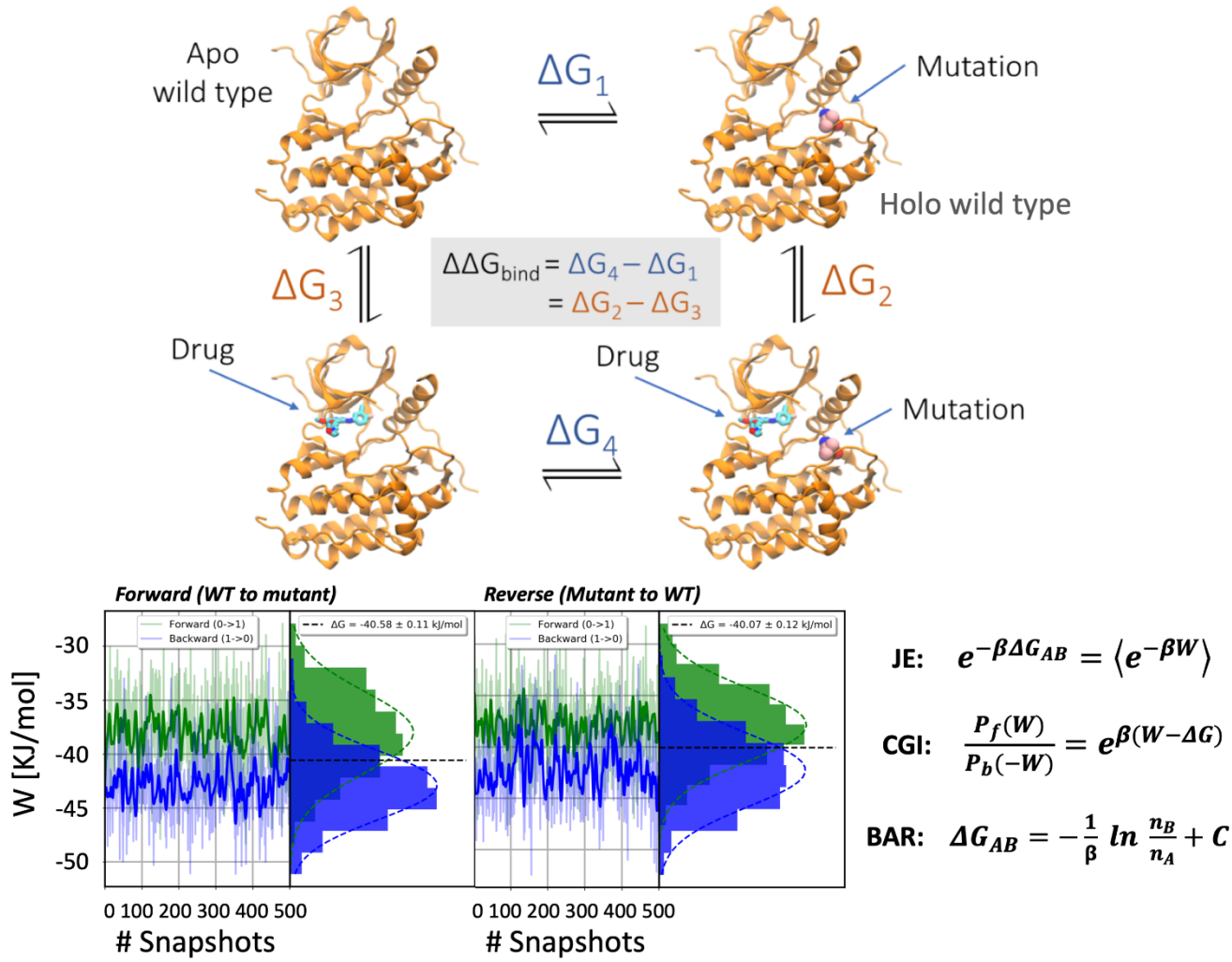


Figure 2: *TOP: Thermodynamic cycle used to determine changes in binding free energy associated with protein mutations. BOTTOM: Examples of histograms of works obtained by mutating one residue into another in apo (left) and holo (right) EGFR, respectively. Equations on the right correspond to the three methods considered here to derive free energies for the histograms of irreversible work (Jarzinski equality (JE; top), Crooks Gaussian Intersection (CGI; middle) and Bennett Acceptance Ratio (BAR; bottom). W stands for the reversible work associated with the A → B mutation, P refers to the histograms (forward and reverse), and C is a constant defined from the A and B partition functions (see Methods).*

Free energy workflows: The calculations above imply the application of a myriad of tools. To use them efficiently, they were organized in execution pipelines (workflows) assembled using the BioExcel Building Blocks library [50] (abbreviated from here onwards as *BioBB*; https://mmb.irbbarcelona.org/biobb/; https://github.com/bioexcel/biobb). A new HPC-focused workflow was specifically developed for this project, handling all the mutations, MD setup and PMX calculations, thereby circumventing human intervention to prepare many thousands of individual simulations (see Figure 3). To ensure efficient usage of pre-exascale computational resources, the workflow was launched and controlled using the PyCOMPSs programming model [83], which automatically distributed the pipeline individual tasks in a parallel manner in HPC supercomputers. A typical run implies the use of c.a. 768 to 3,072 cores of the MareNostrum supercomputer at the Barcelona Supercomputing Center. The protocol has been tested in PRACE supercomputers, showing excellent parallelism on more than 40,000 cores.

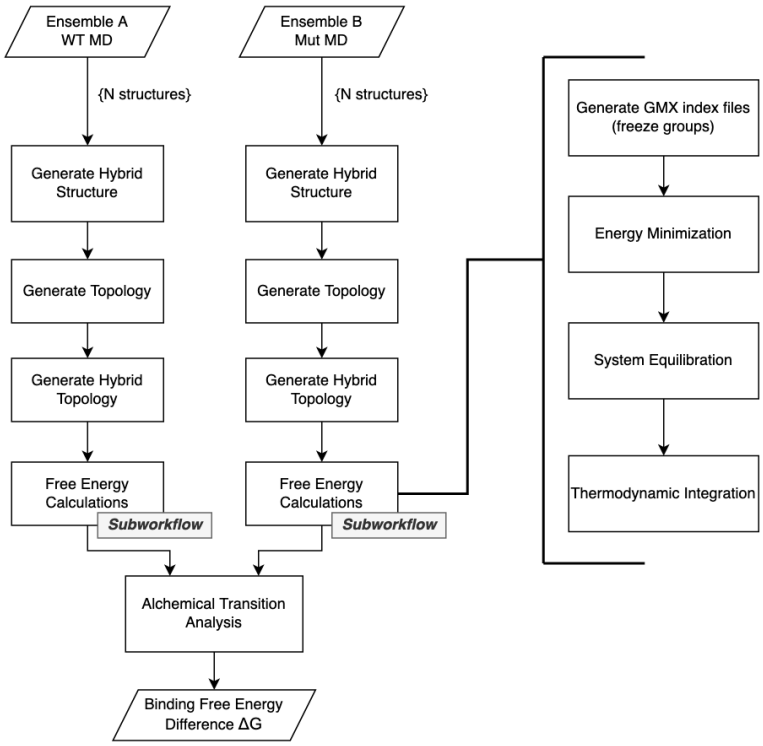


Figure 3: *Relative free energy calculation workflow using BioExcel Building Blocks, wrapping GROMACS and the PMX software. An extended figure can be found in the supplementary material (Suppl. Figure S8).*

RESULTS

Sequence analysis: The ClustalW multiple sequence alignment (Figure 4A) suggests that drug-related mutations tend to be placed at conserved regions (9 out of 14 mutations sites map on regions with Shannon's score H<2), in some cases in ultra-conserved positions such as 796 or 835. However, there are drug-affecting mutations mapping on variable positions and there are many highly conserved positions for which no such mutations are described. Based on sequence conservation, it is difficult to distinguish between positions where resistance-mutations are detected and those where a mutation induces an equal or better drug response. Therefore, multiple alignment analyses do not have enough predictive power to determine which positions are most likely to concentrate mutations affecting drug activity. The analysis of BLOSUM62 matrices [45] suggests that, in general, drug-affecting mutations imply moderate changes in the nature of the amino acid, and no disruptive mutation is detected in the list of drug-affecting mutations (Figure 4B). Interestingly, there is no relationship between the dramatic change induced by a mutation and its impact on drug-resistance. For example, the 'disruptive' change G719S (see Figure 4B) does not lead to resistance, and the 'mild' T790M one inactivates most of the TKIs. Inspection of BLOSUM62 matrices does not allow us to predict drug-affecting mutations. Except for one, all drug-affecting residue changes can be explained by single nucleotide changes, which suggest that they appear spontaneously as human polymorphisms in the absence of drug pressure. However, this is not the case, as shown in Figure 4C. Thus, we can conclude that drug-affecting mutations are a consequence of stressed replication in cancer, which helps to accumulate mutations. Some of these mutations, but not all, will show positive selection as they inactivate response to the drug, thereby enhancing the survival of cancer cells.

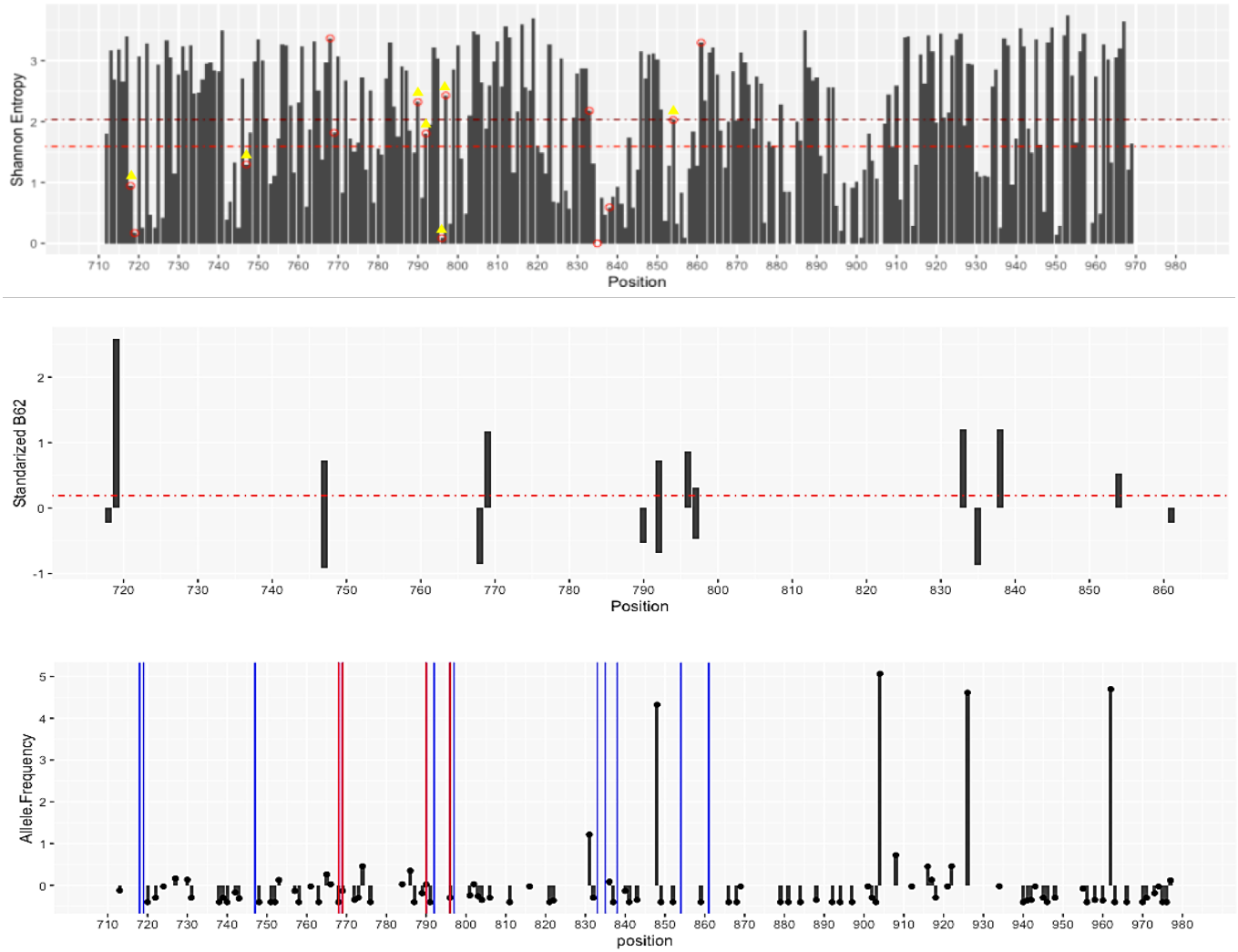


Figure 4. *TOP Shannon's entropy on the human kinase domains (0 fully conserved, 4 fully variable). Places of a drug-affecting mutations are marked with circles and those leading to drug resistance are marked with yellow triangles; the horizontal lines correspond to average Shannon's entropy (see Methods) in the human Tyrosine Kinome for all positions (magenta) and those where drug-affecting mutations are found (red). MIDDLE: Standardized BLOSUM62 index (referred to the expected ones for random mutation of the wild-type residue at this position) associated to the drug-driven mutations, negative values implies that drug-affecting mutations are more aggressive than expected, and positive values the opposite. BOTTOM: Standardized allelic frequencies of polymorphisms in the TK domain of EGFR found in GNOMAD database of human polymorphism. Vertical lines refer to position of drug-affecting mutations (red when mutation is found as a natural polymorphism; blue: drug-affecting mutation does not map a known human polymorphism.*

To complete the sequence analysis, we used PMUT (see Methods) to determine the general pathological potential of positions concentrating drug-affecting mutations, as well as the specific pathological nature of each drug-affecting mutation. Figure 5A shows that, in general, the TK domain (712-979) is the part of EGFR where a higher profile of pathogenicity is expected from any kind of mutations. This observation contrasts with the highly permissive N- and C-terminal domains. In general, positions concentrating drug-affecting mutations are signalled as 'pathological positions' (Figure 5B), but there is no dramatic difference between the average pathogenicity score of drug-affecting positions and the rest of the TK domain (Figure 5B). Finally (Figure 5C), with one exception (the gatekeeper mutation T790M, corresponding to a polymorphism), the rest of the drug-associated mutations imply a pathogenic risk similar to that of a random mutation mapping the same region. This stands for both, mutations leading to an equal or better response of the drug, and those inactivating them. In summary, pathological predictions give almost no clue on whether a mutation should have any impact on the activity of the TKIs.

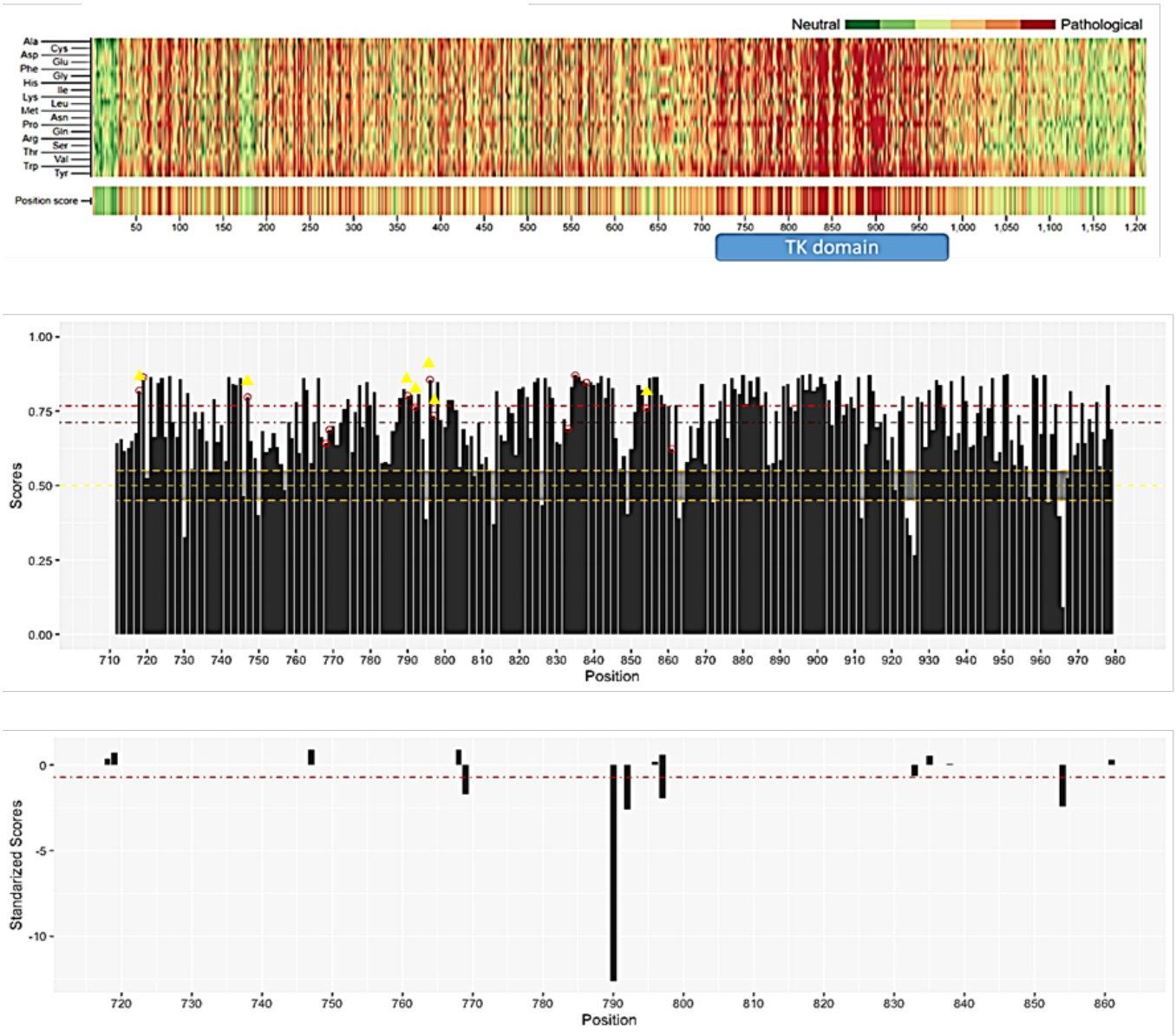


Figure 5: *TOP: Pathogenicity map of EGFR according to PMUT calculations; the upper part of the plot considers the 19 unique mutations at each position and the bottom part the average pathological index at this position (the region of the tyrosine kinase domain is highlighted). MIDDLE: Average pathogenic profile of the tyrosine kinase domain of EGFR (averaging over the 19 potential mutations). The positions of drug-affecting mutations are marked with small circles and mutations generating drug-resistance with yellow triangles; the yellow dashed line indicates the criteria to classify pathological (above) or neutral (below) mutations; the magenta dashed line indicates the average pathogenicity index of the tyrosine kinase domain, and the red one that of the positions where drug-affecting mutations are detected. BOTTOM: Standardized pathogenicity index of the drug-affecting mutations, positive values indicating more pathological than expected and negative values a more neutral effect than expected.*

Interaction energy profiles: As sequence-based techniques fail to have predictive power on the impact of mutations on the modulation of the therapeutic response of drugs, we explore the drug-protein interaction profiles (see Methods) obtained from the ensembles collected from MD simulations. Results in Figure 6 provide interesting information on which residues are dominant in ligand-protein interactions. Thus, for the natural substrate ATP, residues involved in ATP or Mg^{2+} binding are the most prevalent for modulating binding (Figure 6 top left panel). They are: i) highly conserved in multiple TK alignments; ii) rarely involved in polymorphisms; and iii) never mutated due to drug treatment. Therefore, the pressure to have a functional protein protects these crucial residues from mutations. The interaction profiles between protein and inhibitors are quite different to that of the ATP-complex, but the similarity among them is quite surprising (Figure 6),

thereby indicating that all the inhibitors, including third-generation ones, explore a similar region of the protein (note that here we explore the previous non-covalent binding before bond formation). However, there are quantitative differences between the different drugs, especially in terms of the intensity of the interactions with diverse residues. For example, T790 is consistently a stabilizing residue for all inhibitor binding, except Osimertinib, where its effect is negligible, and L718, whose interaction with Osimertinib is much more favourable than with first- and second-generation inhibitors. Globally (Figure 6), around 80% of those mutations leading to drug-resistance occur at positions where WT residues show favourable drug interactions, while around 80% of the mutations leading to no change or even improvement in drug activity are also found at positions where the WT has neutral or unfavourable interactions with the drug. However, drug-affecting mutations are rarely located at residues showing very strong interactions with the drug (labelled in black in Figure 6). Thus, interaction profiles provide information on the regions that are prone to concentrating drug-affecting mutations but are unable to precisely predict the position that can mutate. The interaction profiles when using the experimental structures of the corresponding protein-drug complexes rather than our computational model showed results that are overall similar. This observation is to be expected given the binding site similarities (Suppl. Fig. S9).

To determine whether the interaction profiles can predict the impact of the specific mutation (at a given position) on drug activity, we compare MD-derived interaction profiles from WT and mutant proteins. In general, differential interaction profiles indicate that the protein accommodates the drug-induced mutations well, and the drug-protein interactions outline is not much altered (see Table 1 and examples in Suppl. Figure S10). This result agrees qualitatively with results from the BLOSUM analysis above, confirming that drug-affecting mutations are generally mild. Unfortunately, differential interaction profiles fail to detect some well-characterized resistance mutations, for example, those linked to mutations at L747 and T790. In the first case, the interaction profiles do not recognize L747 as a crucial position for stabilizing the drug (Figure 6), and accordingly, mutations to different residues (Ser, Phe or His) are predicted to be innocuous for binding (see example in Suppl. Figure S10). The case of the T790 gatekeeper is different. Here the energy profile detects threonine as stabilizing (Figure 6), but the substitution to methionine is not predicted to make this interaction weaker. Therefore, some mutations that reduce drug activity do not dramatically alter the direct inhibitor-protein interactions, and the destabilizing effect is expected to be related to structural distortion, solvent bridges, or other effects, which are not captured by these simple calculations.

Exploration of docking landscape: The results above suggest that the impact of some mutations on the binding of inhibitors to the TK domain of EGFR cannot be explained by changes in direct interactions between the drug and the protein. In fact, in some cases, the mutations appear at positions that are clearly involved in inhibitor recognition. Therefore, we explored the docking landscape using PELE (see above), which should allow us to detect changes in binding related not only to direct binding, but also to the easiness of drug entry, or the cost of reorganizing the protein residues at the binding site-aspects that are not considered in simple energy interaction plots. The results in Table 1 show that PELE calculations succeed in predicting drug-induced resistance in 76% of the cases (13 out of 17) compared to 65% of differential interaction energy profiles (17 out of 26). PELE predictions are in general incorrect when water-mediated interactions between the protein and the drug are overlooked (Suppl. Figure S11). These interactions are not explicitly considered in PELE calculations and they stabilize the interactions of the inhibitors with T790 and T854. Exploration of GB/SA estimates (obtained as the total energy of the drug-protein complex minus the energy of the apo-protein) fails to detect any significant trend as the noise in the estimates largely exceed the magnitude to be determined (data not show).

MMPBSA approaches. We tested the ability of MMPBSA calculations to determine the impact of mutation in ligand free energy binding for some pilot systems. As expected, results were not robust as binding free energies are obtained by subtracting two very large numbers, which leads to large errors. When these estimates

are combined to obtain mutation-induced changes in binding free energy, the accumulated errors are even larger, leading to big uncertainties and no predictive power (see pilot calculations in Supp. Table S1).

ML approaches. We tested the ability of ML approach to predict drug-affecting mutation (see Methods). Unfortunately, in our hands poor results were found and the Random Forest-based PremPLI method, which takes structural details into account, does not have any predictive power, as it shows a marked tendency towards considering all the mutations 'resistance-driven' (Suppl. Table S2).

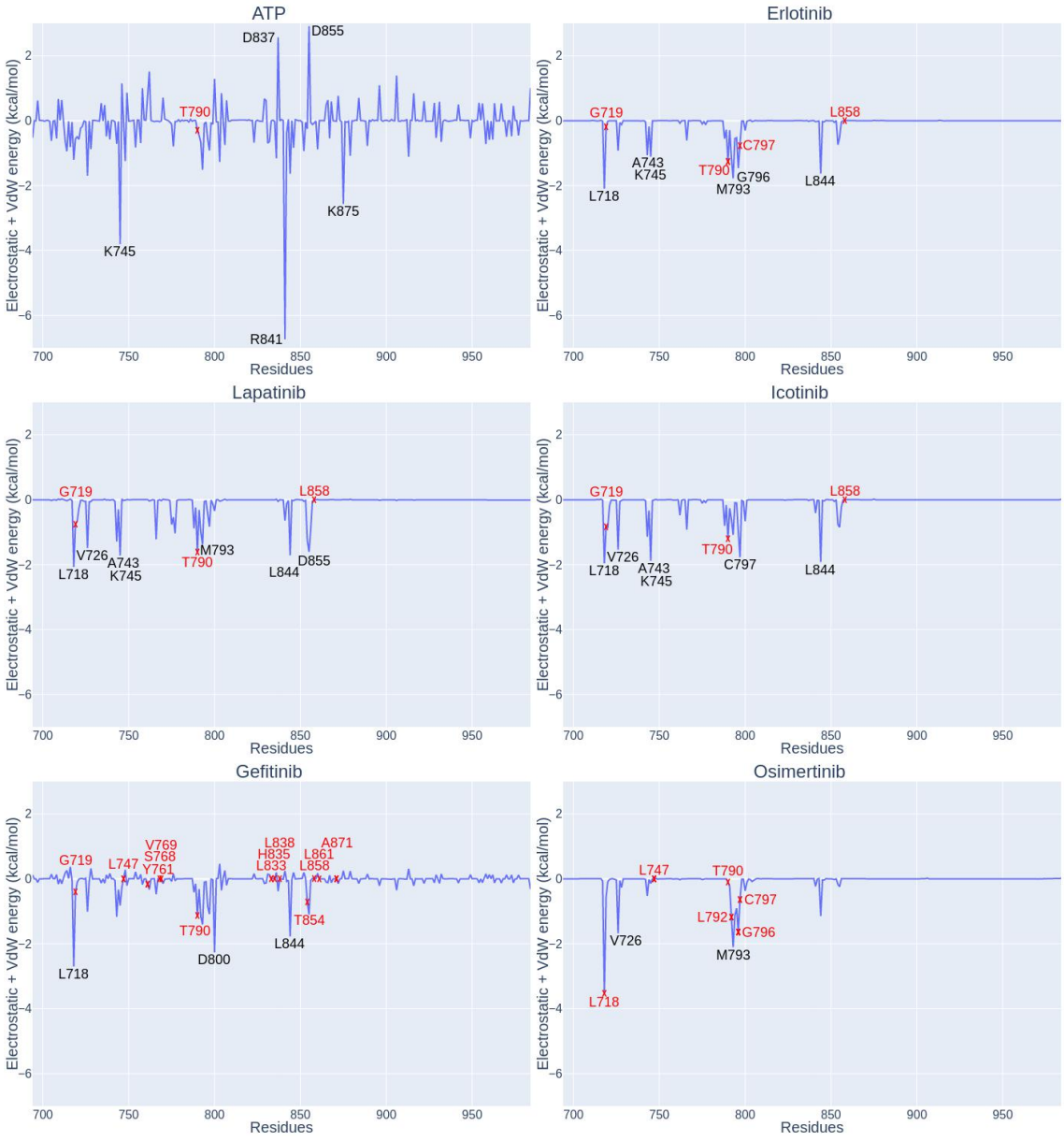


Figure 6: *Interaction profiles for natural substrate ATP and inhibitors of the tyrosine kinase domain of EGFR. Residues involved in the interaction with the binder are shown in black and mutations that affect binding in red.*

Molecular dynamics-based free energy calculations: These calculations combine equilibrium MD simulations with non-equilibrium alchemical mutations in the apo and holo states of the protein and represent the 'state-of-the-art' in atomistic simulations. Traditionally, its practical use requires thorough expertise and considerable computational resources, but the workflow developed here (Figure 3) allows automation, full use of massively parallel computer architectures, and simple use even for non-experts. The only source of uncertainty arises from the small divergence in the estimates among different integration methods. The analysis of histograms (Figure 2) reveals that problematic cases are typically related to poor overlap between the 'forward' and the 'reverse' histograms, which generate noise that is maximized in CGI where a Gaussian

distribution of irreversible work is assumed. Very encouragingly (Table 1), these divergences are detected in only a few cases and are corrected by simply extending simulations. By construction, FE/MD methods contain all the enthalpic and entropic contributions to binding; this combined with a good force-field, and appropriate simulation lengths, should provide accurate estimates. Indeed, the predictive power of the FE/MD protocol outlined here is ideal, as it succeeds in correctly classifying mutations in all the studied cases, even in those where simpler methods based on the analysis of differential energy profiles or Monte Carlo PELE calculations fail. We cannot expect this performance to translate to all proteins, drugs, and mutations, but the excellent results obtained on a related system also using alchemical free energy simulations [84] raises optimism that these sophisticated simulations might be of general use to predict the impact of single point mutations on drug activity, even at preclinical or early clinical stages. The intrinsic complexity of these calculations that limits their use to a small number of highly expert groups is reduced by the development of robust workflows, whose use does not require expertise and allows results to be obtained at a time scale compatible with pre-clinical and clinical use.

Mutation	**Drug**	**Eprof (pred)**	**PELE* (pred)**	**ΔΔG(binding) FE/MD**	**Exp. Impact$^{\Delta}$**
L718Q	Osimertinib	R	-	18.7(0.8) R	Resistance[1]
G719S	Gefitinib	S	S	-5.5(0.9) S	Sensitive[2]
G719S	Icotinib	S	-	-0.9(0.8)# S	Sensitive[2]
G719S	Erlotinib	S	S	-1.1(0.3) S	Sensitive[3]
G719S	Lapatinib	S	S	-5.1(2.8) S	Sensitive[4]
L747S	Gefitinib	*S*	*S*	13.8(1.8) R	Resistance[5]
L747F	Osimertinib	*S*	-	4.4(1.2) R	Resistance[6]
L747H	Osimertinib	*S*	-	6.5(3.6) R	Resistance[6]
S768I	Gefitinib	S	*R*	-1.8(0.4) S	Sensitive[7]
V769M	Gefitinib	S	S	-7.7(2.2) S	Sensitive[8]
T790M	Gefitinib	*S*	*S*	15.5(0.9) R	Resistance[9]
T790M	Erlotinib	*S*	R	17.6(1.9) R	Resistance[9]
T790M	Lapatinib	*S*	R	19.3(0.9) R	Resistance[10]
T790M	Osimertinib	S	-	-0.5(2.1) S	Sensitive[11]
T790M	Icotinib	S	-	11.5(0.6) R	Resistance[12]
L792F	Osimertinib	S	-	4.2(1.0) R	Resistance[13]
L792H	Osimertinib	R	-	8.8(0.1) R	Resistance[13]
G796S	Osimertinib	*S*	-	4.8(0.4) R	Resistance[14]
C797G&	Osimertinib	R	R	Resistance	Resistance[15]
C797S&	Osimertinib	R	R	Resistance	Resistance[16]
L833V	Gefitinib	S	S	-5.7(1.4) S	Sensitive[17]
H835L	Gefitinib	S	S	-7.5(1.3)# S	Sensitive[17]
L838V	Gefitinib	S	S	-6.3(3.1) S	Sensitive[18]
T854A	Gefitinib	R	S	13.2(1.1) R	Resistance[5]
L861Q	Gefitinib	S	S	-5.4(0.7)# S	Sensitive[19]
T790M/C797S	Erlotinib	R	R	6.2(2.4) R	Resistance[20]

Table 1. Mutations impacting drug activity and estimates of the effect based on interaction energy profiles (3rd column), PELE docking (4th column), and molecular dynamics-based free energy simulations (5th column; with standard deviations on the values shown in brackets); for the latter, the predicted change in free energy of binding is included in kJ/mol. The experimental annotation of the effect of the mutation on drug activity is shown in the last column. In all cases, grey cells indicate prediction errors.

* PELE calculations require X-ray structures as an input and cannot be used to explore covalent binders.

& A trivial prediction of resistance for covalent inhibitors binding to C797

Due to poor convergence on 100*2*2 histograms, the calculations were extended to 500*2*2 TI trajectories.

[Δ] References for experimental activities are displayed in the Supplementary Material.

DISCUSSION

The mechanisms behind mutation-induced drug resistance are diverse, even in those cases where mutations map on the drug-binding site. Thus, mutations might affect the stability of the protein, protein-protein interactions, the resistance of proteins to degradation, the entry of the drug, or the inactive/active conformational equilibrium [85, 86] thereby increasing the 'active state' and accordingly the affinity for the substrate, which could mask the inhibitory properties of the drug. For example, in the case of EFGR, the T790M mutation has been suggested to increase the affinity for ATP by displacing the equilibrium towards the 'active state' [12, 87]. However, T790M is not reported as a 'cancer-driven' mutation, which would be expected for mutations inducing constitutive kinase activity. Furthermore, most of the inhibitors bind to the 'active state', which means that the impact of inactive/active conformational transition in terms of ATP vs. drug binding should be not dramatic. The requirement for EGFR to maintain the kinase activity rules out resistance-related mutations affecting the global structure or the entry of the substrate/drug. Similar reasonings are likely to be transferable to other proteins, thus suggesting that interfering with drug binding is likely to be a common mechanism in resistance-related mutations [76-78], even though other processes can eventually contribute to the resistance. We assume here that we can predict the sensitive/resistance nature of a given mutation (for a particular drug) based on the fingerprint that it produces in the binding free energy of the drug. The question is then how to obtain an estimate of the impact of mutations on drug binding compatible with the needs of clinical practice or preclinical research.

Sequence analyses provide useful information on the origin and placement of drug-affecting mutations. In most cases, these alterations are generated by single nucleotide changes and are typically located in conserved regions, where the pathogenic risk associated with mutations is high. Specific positions where drug-affecting mutations occur show a similar degree of conservation to that of neighbouring regions. The mutations that lead to an alteration in drug efficacy tend to be mild in terms of changes in amino acid properties and are not more pathogenic than the average expected value at that position. With a few exceptions, drug-associated mutations do not match polymorphisms. This observation suggests that high stress in replication and most likely poor proofreading of the nascent DNA are required for the appearance of these mutations. Finally, a significant number of the studied mutations imply equal or higher sensitivity to the drug than the WT, which means that not all drug-affecting mutations correspond to a canonical positive selection paradigm. Overall, sequence-dependent trends are useful to define regions where mutations are susceptible to altering the response to the drug, but are not able to predict when a mutation will cause resistance to chemotherapy.

Energy profiles efficiently detect those regions that establish strong interactions with the ligand and, accordingly, are more informative than sequence analysis to precisely detect the 'susceptible' regions where mutations might impact the activity of the drug. However, the success rate in predicting drug-affecting mutations is only moderate as there is a non-negligible number of cases where the impact of mutations on ligand binding is modulated by non-direct interacting terms. When flexibility and diffusion considerations are incorporated in the evaluation of drug binding, the predictive power increases, but not dramatically (up to 76%), with cases where we cannot reproduce experimental findings, in most cases due to the involvement of water-mediated interactions that are not easily captured by a method based on continuum solvation models. GB/SA calculations have large associated errors, which hamper any meaningful comparison and, in our hands,

they lack predictive power. Similarly, although ML-based approaches rely on structural information and are specifically trained to predict the impact of mutations on drug binding, they show very poor predictive power.

Non-equilibrium alchemical free energy calculations provide results of an astonishing accuracy (100% success rate), based only on physical principles without any *ad hoc* training process. By construction, assuming a good force field and extended sampling, the protocol should capture the different contributions (enthalpic and entropic) to differential binding and has the advantage that it provides a physical rationale for the effect of the studied mutations. The limitations of these types of calculations are clear: i) they require user expertise; ii) the setup of the calculations is difficult as it involves thousands of individual simulations, each requiring several preparation steps; and iii) these calculations are computationally expensive and might require very large wall clock times, thereby hampering its practical use in clinical environments. The BioExcel Building Blocks-based workflow developed here allows us to greatly simplify the complexity of launching simulations, thereby circumventing the need for specific training in advanced simulation methods. Furthermore, the use of a clever workflow manager (PyCOMPSs; see Methods) allows extremely fast and efficient parallelism, thereby reducing the entire process to hours when using a medium-sized cluster. It can reduce the process to minutes in a pre-exascale supercomputer and most likely to seconds in an ExaScale machine. We speculate that once fully calibrated and tested, protocols like the one shown here could be used to accurately predict mutations affecting drug activity in the *in silico* stages of drug design, thereby contributing to the development of alternative drugs by anticipating inactivating mutations.

DATA AND SOFTWARE AVAILABILITY

The code for this work is available at https://github.com/bioexcel/biobb_hpc_workflows/tree/condapack. MD simulations are available from our MDposit repository: https://mdposit-dev.bsc.es/#/browse?search=egfr

SUPPORTING INFORMATION AVAILABILITY

Supporting Information Available: Supplementary Figures (S1 to S11) and Supplementary Tables S1 and S2 (with references).

ACKNOWLEDGEMENTS

We are indebted to the BioExcel partners, especially Prof. de Groot's group for helpful discussion on PMX calculations and Prof. Rosa Mª Badia for help with the PyCOMPSs programming model. This work has been supported by the BioExcel-2: Centre of Excellence for Computational Biomolecular Research (823830), the Spanish Ministry of Science (RTI2018-096704-B-100, PID2020-116620GB-I00), and the Instituto de Salud Carlos III–Instituto Nacional de Bioinformática (ISCIII PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional). Funding was also provided by the MINECO Severo Ochoa Award of Excellence from the Government of Spain (awarded to IRB Barcelona). M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avancats) Academia researcher. Nostrum Biodiscovery is supported by the Fundación Marcelino Botín (Mind the Gap), CDTI (Neotec grant -EXP 00094141/SNEO-20161127), and a Torres Quevedo grant (PTQ2018-009992).

REFERENCES

(1) Hanahan, D.; Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144* (5), 646-674. DOI: 10.1016/j.cell.2011.02.013.
(2) Hanahan, D.; Weinberg, R. A. The hallmarks of cancer. *Cell* **2000**, *100* (1), 57-70. DOI: 10.1016/s0092-8674(00)81683-9.

(3) Hirsch, F. R.; Varella-Garcia, M.; Bunn, P. A.; Di Maria, M. V.; Veve, R.; Bremmes, R. M.; Barón, A. E.; Zeng, C.; Franklin, W. A. Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. *J Clin Oncol* **2003**, *21* (20), 3798-3807. DOI: 10.1200/JCO.2003.11.069.
(4) Sharma, S. V.; Bell, D. W.; Settleman, J.; Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer* **2007**, *7* (3), 169-181. DOI: 10.1038/nrc2088.
(5) Chen, Y. M. Update of epidermal growth factor receptor-tyrosine kinase inhibitors in non-small-cell lung cancer. *J Chin Med Assoc* **2013**, *76* (5), 249-257. DOI: 10.1016/j.jcma.2013.01.010.
(6) Kris, M. G.; Johnson, B. E.; Kwiatkowski, D. J.; Iafrate, A. J.; Wistuba, I. I.; Aronson, S. L.; Engelman, J. A.; Shyr, Y.; Khuri, F. R.; Rudin, C. M.; Garon, E. B.; Pao, W; Schiller, J. H.; Haura, E. B.; Shirai, K.; Giaccone, G.; Berry, L. D.; Kugler, K.; Minna, J. D.; Bunn, P. A. Identification of driver mutations in tumor specimens from 1,000 patients with lung adenocarcinoma: The NCI's Lung Cancer Mutation Consortium (LCMC). *Journal of Clinical Oncology* **2011**, *29* (18_suppl), CRA7506-CRA7506. DOI: 10.1200/jco.2011.29.18_suppl.cra7506 (acccessed 2022/10/06).
(7) Shiau, C. J.; Babwah, J. P.; da Cunha Santos, G.; Sykes, J. R.; Boerner, S. L.; Geddie, W. R.; Leighl, N. B.; Wei, C.; Kamel-Reid, S.; Hwang, D. M.; Ming-Sound T. Sample features associated with success rates in population-based EGFR mutation testing. *J Thorac Oncol* **2014**, *9* (7), 947-956. DOI: 10.1097/JTO.0000000000000196.
(8) Orellana, L.; Thorne, A. H.; Lema, R.; Gustavsson, J.; Parisian, A. D.; Hospital, A.; Cordeiro, T. N.; Bernadó, P.; Scott, A. M.; Brun-Heath, I.; Lindahl, E.; Cavenee, W. K.; Furnari, F. B.; Orozco, M. Oncogenic mutations at the EGFR ectodomain structurally converge to remove a steric hindrance on a kinase-coupled cryptic epitope. *Proc Natl Acad Sci U S A* **2019**, *116* (20), 10009-10018. DOI: 10.1073/pnas.1821442116.
(9) Choi, S. H.; Mendrola, J. M.; Lemmon, M. A. EGF-independent activation of cell-surface EGF receptors harboring mutations found in gefitinib-sensitive lung cancer. *Oncogene* **2007**, *26* (11), 1567-1576. DOI: 10.1038/sj.onc.1209957.
(10) Kumar, A.; Petri, E. T.; Halmos, B.; Boggon, T. J. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. *J Clin Oncol* **2008**, *26* (10), 1742-1751. DOI: 10.1200/JCO.2007.12.1178.
(11) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer T. M.; Shewchuk L. A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res* **2004**, *64* (18), 6652-6659. DOI: 10.1158/0008-5472.CAN-04-1168.
(12) Yun, C. H.; Mengwasser, K. E.; Toms, A. V.; Woo, M. S.; Greulich, H.; Wong, K. K.; Meyerson, M.; Eck, M. J. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc Natl Acad Sci U S A* **2008**, *105* (6), 2070-2075. DOI: 10.1073/pnas.0709662105.
(13) Zhang, X.; Gureasko, J.; Shen, K.; Cole, P. A.; Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **2006**, *125* (6), 1137-1149. DOI: 10.1016/j.cell.2006.05.013.
(14) Ruan, Z.; Kannan, N. Altered conformational landscape and dimerization dependency underpins the activation of EGFR by αC–β4 loop insertion mutations. *Proceedings of the National Academy of Sciences* **2018**, *115* (35), E8162-E8171. DOI: 10.1073/pnas.1803152115 (acccessed 2022/10/06).
(15) Wan, S.; Coveney, P. V. Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. *J Comput Chem* **2011**, *32* (13), 2843-2852. DOI: 10.1002/jcc.21866.
(16) Rocha-Lima, C. M.; Soares, H. P.; Raez, L. E.; Singal, R. EGFR targeting of solid tumors. *Cancer Control* **2007**, *14* (3), 295-304. DOI: 10.1177/107327480701400313.
(17) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J Biol Chem* **2002**, *277* (48), 46265-46272. DOI: 10.1074/jbc.M207135200.
(18) Pao, W.; Chmielecki, J. Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. *Nat Rev Cancer* **2010**, *10* (11), 760-774. DOI: 10.1038/nrc2947.

(19) Sequist, L. V.; Joshi, V. A.; Jänne, P. A.; Muzikansky, A.; Fidias, P.; Meyerson, M.; Haber, D. A.; Kucherlapati, R.; Johnson, B. E.; Lynch, T. J. Response to treatment and survival of patients with non-small cell lung cancer undergoing somatic EGFR mutation testing. *Oncologist* **2007**, *12* (1), 90-98. DOI: 10.1634/theoncologist.12-1-90.
(20) Stella, G. M.; Luisetti, M.; Inghilleri, S.; Cemmi, F.; Scabini, R.; Zorzetto, M.; Pozzi, E. Targeting EGFR in non-small-cell lung cancer: lessons, experiences, strategies. *Respir Med* **2012**, *106* (2), 173-183. DOI: 10.1016/j.rmed.2011.10.015.
(21) Wakeling, A. E.; Guy, S. P.; Woodburn, J. R.; Ashton, S. E.; Curry, B. J.; Barker, A. J.; Gibson, K. H. ZD1839 (Iressa): an orally active inhibitor of epidermal growth factor signaling with potential for cancer therapy. *Cancer Res* **2002**, *62* (20), 5749-5754.
(22) Cross, D. A.; Ashton, S. E.; Ghiorghiu, S.; Eberlein, C.; Nebhan, C. A.; Spitzler, P. J.; Orme, J. P.; Finlay, M. R.; Ward, R. A.; Mellor, M. J.; Hughes, G.; Rahi, A.; Jacobs, V. N.; Brewer, M. R.; Ichihara, E.; Sun, J.; Jin, H.; Ballard, P.; Al-Kadhimi, K.; Rowlinson, R.; Klinowska, T.; Richmond, G. H. P.; Cantarini, M.; Kim, D.; Ranson, M. R.; Pao, W. AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov* **2014**, *4* (9), 1046-1061. DOI: 10.1158/2159-8290.CD-14-0337.
(23) Wang, S.; Cang, S.; Liu, D. Third-generation inhibitors targeting EGFR T790M mutation in advanced non-small cell lung cancer. *J Hematol Oncol* **2016**, *9*, 34. DOI: 10.1186/s13045-016-0268-z.
(24) Huang, L.; Fu, L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. *Acta Pharm Sin B* **2015**, *5* (5), 390-401. DOI: 10.1016/j.apsb.2015.07.001.
(25) Becker, A.; Crombag, L.; Heideman, D. A.; Thunnissen, F. B.; van Wijk, A. W.; Postmus, P. E.; Smit, E. F. Retreatment with erlotinib: Regain of TKI sensitivity following a drug holiday for patients with NSCLC who initially responded to EGFR-TKI treatment. *Eur J Cancer* **2011**, *47* (17), 2603-2606. DOI: 10.1016/j.ejca.2011.06.046.
(26) Jang, S. H. Long Term Therapeutic Plan for Patients with Non-Small Cell Lung Cancer Harboring EGFR Mutation. *Tuberc Respir Dis (Seoul)* **2014**, *76* (1), 8-14. DOI: 10.4046/trd.2014.76.1.8.
(27) Oxnard, G. R.; Janjigian, Y. Y.; Arcila, M. E.; Sima, C. S.; Kass, S. L.; Riely, G. J.; Pao, W.; Kris, M. G.; Ladanyi, M.; Azzoli, C. G.; Miller, V. A. Maintained Sensitivity to EGFR Tyrosine Kinase Inhibitors in EGFR-Mutant Lung Cancer Recurring after Adjuvant Erlotinib or Gefitinib. *Clinical Cancer Research* **2011**, *17* (19), 6322-6328. DOI: 10.1158/1078-0432.CCR-11-1080 (acccessed 10/6/2022).
(28) Song, T.; Yu, W.; Wu, S. X. Subsequent Treatment Choices for Patients with Acquired Resistance to EGFR-TKIs in Non-small Cell Lung Cancer: Restore after a Drug Holiday or Switch to another EGFR-TKI? *Asian Pacific Journal of Cancer Prevention* **2014**, *15* (1), 205-213.
(29) Song, Z.; Yu, X.; He, C.; Zhang, B.; Zhang, Y. Re-administration after the failure of gefitinib or erlotinib in patients with advanced non-small cell lung cancer. *Journal of Thoracic Disease* **2013**, *5* (4), 400-405.
(30) Stewart, E. L.; Tan, S. Z.; Liu, G.; Tsao, M. S. Known and putative mechanisms of resistance to EGFR targeted therapies in NSCLC patients with EGFR mutations-a review. *Transl Lung Cancer Res* **2015**, *4* (1), 67-81. DOI: 10.3978/j.issn.2218-6751.2014.11.06.
(31) Kancha, R. K.; von Bubnoff, N.; Peschel, C.; Duyster, J. Functional analysis of epidermal growth factor receptor (EGFR) mutations and potential implications for EGFR targeted therapy. *Clin Cancer Res* **2009**, *15* (2), 460-467. DOI: 10.1158/1078-0432.CCR-08-1757.
(32) Jaeger, S.; Igea, A.; Arroyo, R.; Alcalde, V.; Canovas, B.; Orozco, M.; Nebreda, A. R.; Aloy, P. Quantification of Pathway Cross-talk Reveals Novel Synergistic Drug Combinations for Breast Cancer. *Cancer Res* **2017**, *77* (2), 459-469. DOI: 10.1158/0008-5472.CAN-16-0097.
(33) Juchum, M.; Günther, M.; Laufer, S. A. Fighting cancer drug resistance: Opportunities and challenges for mutation-specific EGFR inhibitors. *Drug Resist Updat* **2015**, *20*, 12-28. DOI: 10.1016/j.drup.2015.05.002.
(34) Tan, C. S.; Kumarakulasinghe, N. B.; Huang, Y. Q.; Ang, Y. L. E.; Choo, J. R.; Goh, B. C.; Soo, R. A. Third generation EGFR TKIs: current data and future directions. *Mol Cancer* **2018**, *17* (1), 29. DOI: 10.1186/s12943-018-0778-0.
(35) Zehir, A.; Benayed, R.; Shah, R. H.; Syed, A.; Middha, S.; Kim, H. R.; Srinivasan, P.; Gao, J.; Chakravarty, D.; Devlin, S. M.; Hellmann, M. D.; Barron, D. A.; Schram, A. M.; Hameed, M.; Dogan, S.; Ross, D. S.; Hechtman, J. F.; DeLair, D. F.; Yao, J.; Mandelker, D. L.; Cheng, D. T.; Chandramohan, R.; Mohanty, A. S.; Ptashkin, R. N.; Jayakumaran, G.; Prasad, M.; Syed, M. H.; Rema, A. B.; Liu, Z. Y.; Nafa,

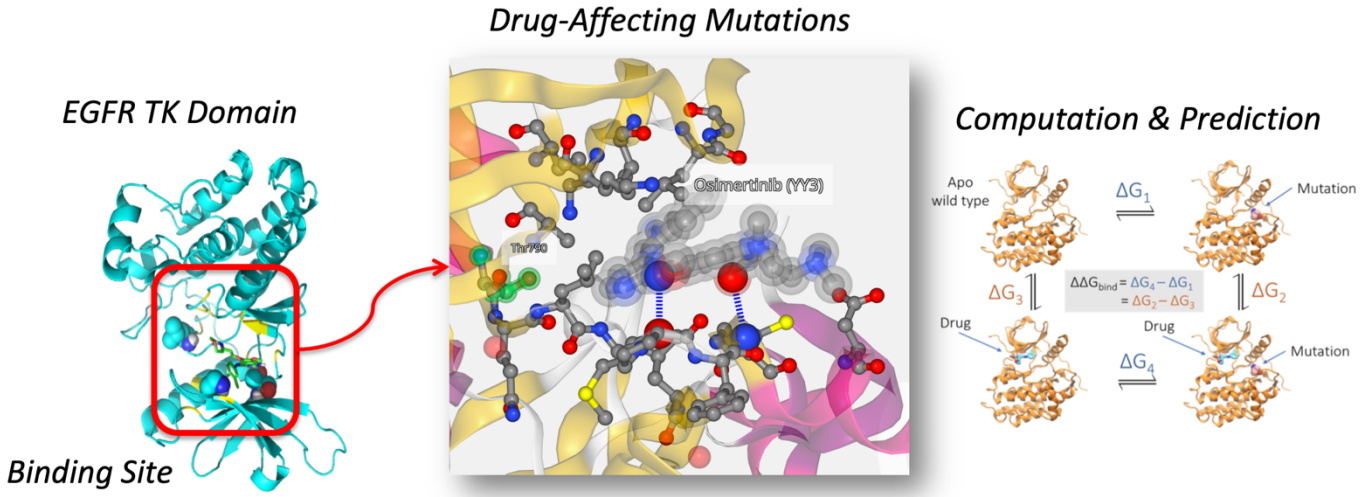
K.; Borsu, L.; Sadowska, J.; Casanova, J.; Bacares, R.; Kiecka, I. J.; Razumova, A.; Son, J. B.; Stewart, L.; Baldi, T.; Mullaney, K. A.; Al-Ahmadie, H.; Vakiani, E.; Abeshouse, A. A.; Penson, A. V.; Jonsson, P.; Camacho, N.; Chang, M. T.; Won, H. H.; Gross, B. E.; Kundra, R.; Heins, Z. J.; Chen, H-W.; Phillips, S.; Zhang, H.; Wang, J.; Ochoa, A.; Wills, J.; Eubank, M.; Thomas, S. B.; Gardos, S. M.; Reales, D. N.; Galle, J.; Durany, R.; Cambria, R.; Abida, W.; Cercek, A.; Feldman, D. R.; Gounder, M. M.; Hakimi, A. A.; Harding, J. J.; Iyer, G.; Janjigian, Y. Y.; Jordan, E. J.; Kelly, C. M.; Lowery, M. A.; Morris, L. G. T.; Omuro, A. M.; Raj, N.; Razavi, P.; Shoushtari, A. N.; Shukla, N.; Soumerai, T. E.; Varghese, A. M.; Yaeger, R.; Coleman, J.; Bochner, B.; Riely, G. J.; Saltz, L. B.; Scher, H. I.; Sabbatini, P. J.; Robson, M. E.; Klimstra, D. S.; Taylor, B. S.; Baselga, J.; Schultz, N.; Hyman, D. M.; Arcila, M. E.; Solit, D. B.; Ladanyi, M.; Berger, M. F. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **2017**, *23* (6), 703-713. DOI: 10.1038/nm.4333.
(36) Hudson, T. J.; Anderson, W.; Artez, A.; Barker, A. D.; Bell, C.; Bernabé, R. R.; Bhan, M. K.; Calvo, F.; The International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **2010**, *464* (7291), 993-998. DOI: 10.1038/nature08987.
(37) Forbes, S. A.; Beare, D.; Gunasekaran, P.; Leung, K.; Bindal, N.; Boutselakis, H.; Ding, M.; Bamford, S.; Cole, C.; Ward, S.; Kok, C. Y.; Jia, M.; De, T.;: Teague, J. W.; Stratton, M. R.; McDermott, U.; Campbell, P. J. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **2015**, *43* (D1), D805-D811. DOI: 10.1093/nar/gku1075 (acccessed 10/6/2022).
(38) Landrum, M. J.; Lee, J. M.; Riley, G. R.; Jang, W.; Rubinstein, W. S.; Church, D. M.; Maglott, D. R. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **2014**, *42* (Database issue), D980-985. DOI: 10.1093/nar/gkt1113.
(39) Manning, G. *KinBase: The Kinase Database*. 2012. http://kinase.com/web/current/kinbase/ (accessed.
(40) Bodenhofer, U.; Bonatesta, E.; Horejš-Kainrath, C.; Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **2015**, *31* (24), 3997-3999. DOI: 10.1093/bioinformatics/btv494.
(41) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **1994**, *22*. DOI: 10.1093/nar/22.22.4673.
(42) Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **2003**, *31* (13), 3497-3500. DOI: 10.1093/nar/gkg500.
(43) Shannon, C. E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press, 1949.
(44) Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alföldi, J.; Wang, Q.; Collins, R. L.; Laricchia, K. M.; Ganna, A.; Birnbaum, D. P.; Gauthier, L. D.; Brand, H.; Solomonson, M.; Watts, N. A.; Rhodes, D.; Singer-Berk, M.; England, E. M.; Seaby, E. G.; Kosmicki, J. A.; Walters, R. K.; Tashman, K.; Farjoun, Y.; Banks, E.; Poterba, T.; Wang, A.; Seed, C.; Whiffin, N.; Chong, J. X.; Samocha, K. E.; Pierce-Hoffman, E.; Zappala, Z.; O'Donnell-Luria, A. H.; Minikel, E. V.; Weisburd, B.; Lek, M.; Ware, J. S.; Vittal, C.; Armean, I. M.; Bergelson, L.; Cibulskis, K.; Connolly, K. M.; Covarrubias, M.; Donnelly, S.; Ferriera, S.; Gabriel, S.; Gentry, J.; Gupta, N.; Jeandet, T.; Kaplan, D.; Llanwarne, C.; Munshi, R.; Novod, S.; Petrillo, N.; Roazen, D.; Ruano-Rubio, V.; Saltzman, A.; Schleicher, M.; Soto, J.; Tibbetts, K.; Tolonen, C.; Wade, G.; Talkowski, M. E.; Genome Aggregation Database Consortium; Neale, B. M.; Daly, M. J.; MacArthur, D. G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581* (7809), 434-443. DOI: 10.1038/s41586-020-2308-7.
(45) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **1992**, *89* (22), 10915-10919. DOI: 10.1073/pnas.89.22.10915.
(46) López-Ferrando, V.; Gazzo, A.; de la Cruz, X.; Orozco, M.; Gelpí, J. L. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research* **2017**, *45* (W1), W222-W228. DOI: 10.1093/nar/gkx313 (acccessed 12/19/2019).
(47) Ferrer-Costa, C.; Gelpí, J. L.; Zamakola, L.; Parraga, I.; de la Cruz, X.; Orozco, M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **2005**, *21* (14), 3176-3178. DOI: 10.1093/bioinformatics/bti486.

(48) Ferrer-Costa, C.; Orozco, M.; de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* **2002**, *315* (4), 771-786. DOI: 10.1006/jmbi.2001.5255.
(49) Ferrer-Costa, C.; Orozco, M.; de la Cruz, X. Sequence-based prediction of pathological mutations. *Proteins* **2004**, *57* (4), 811-819. DOI: 10.1002/prot.20252.
(50) Andrio, P.; Hospital, A.; Conejero, J.; Jordá, L.; Del Pino, M.; Codo, L.; Soiland-Reyes, S.; Goble, C.; Lezzi, D.; Badia, R. M.; Orozco, M.; Gelpí, J. L. BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. *Sci Data* **2019**, *6* (1), 169. DOI: 10.1038/s41597-019-0177-4.
(51) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010**, *31* (2), 455-461. DOI: 10.1002/jcc.21334.
(52) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78* (8), 1950-1958. DOI: 10.1002/prot.22711.
(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935. DOI: doi:http://dx.doi.org/10.1063/1.445869.
(54) Moreno, D.; Zivanovic, S.; Colizzi, F.; Hospital, A.; Aranda, J.; Soliva, R.; Orozco, M. DFFR: A New Method for High-Throughput Recalibration of Automatic Force-Fields for Drugs. *J Chem Theory Comput* **2020**, *16* (10), 6598-6608. DOI: 10.1021/acs.jctc.0c00306.
(55) Zivanovic, S.; Bayarri, G.; Colizzi, F.; Moreno, D.; Gelpí, J. L.; Soliva, R.; Hospital, A.; Orozco, M. Bioactive Conformational Ensemble Server and Database. A Public Framework to Speed Up. *J Chem Theory Comput* **2020**, *16* (10), 6586-6597. DOI: 10.1021/acs.jctc.0c00305.
(56) Zivanovic, S.; Colizzi, F.; Moreno, D.; Hospital, A.; Soliva, R.; Orozco, M. Exploring the Conformational Landscape of Bioactive Small Molecules. *J Chem Theory Comput* **2020**, *16* (10), 6575-6585. DOI: 10.1021/acs.jctc.0c00304.
(57) 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*. DOI: 10.1002/jcc.20035.
(58) Olsson, M. H.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **2011**, *7* (2), 525-537. DOI: 10.1021/ct100578z.
(59) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; Chun, M.; Li, P.; Gohara, D. W.; Dolinsky, T.; Konecny, R.; Koes, D. R.; Nielsen, J. E.; Head-Gordon, T.; Geng, W.; Krasny, R.; Wei, G-W.; Holst, M. J.; McCammon, J. A.; Baker, N. A. Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **2018**, *27* (1), 112-128. DOI: 10.1002/pro.3280.
(60) Borrelli, K. W.; Vitalis, A.; Alcantara, R.; Guallar, V. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *J Chem Theory Comput* **2005**, *1* (6), 1304-1311. DOI: 10.1021/ct0501811.
(61) *Schrödinger Release. Protein preparation wizard.*; 2018. (accessed.
(62) *Schrödinger Release. Maestro.*; 2017. (accessed.
(63) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B* **2001**, *105* (28), 6474-6487. DOI: 10.1021/jp003919d.
(64) Zhou, R.; Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proceedings of the National Academy of Sciences* **2002**, *99* (20), 12777-12782. DOI: doi:10.1073/pnas.142430099.
(65) Lecina, D.; Gilabert, J. F.; Guallar, V. Adaptive simulations, towards interactive protein-ligand modeling. *Sci Rep* **2017**, *7* (1), 8466. DOI: 10.1038/s41598-017-08445-5.
(66) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997**, *18* (12), 1463-1472, https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H. DOI:

https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H (acccessed 2022/10/06).
(67) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98* (12), 10089-10092. DOI: doi:http://dx.doi.org/10.1063/1.464397.
(68) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103* (19), 8577-8593. DOI: 10.1063/1.470117 (acccessed 2022/10/06).
(69) Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **1983**, *50* (5), 1055-1076. DOI: 10.1080/00268978300102851.
(70) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52* (12), 7182-7190. DOI: 10.1063/1.328693 (acccessed 2022/10/06).
(71) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19-25. DOI: http://dx.doi.org/10.1016/j.softx.2015.06.001.
(72) Soliva, R.; Gelpí, J. L.; Almansa, C.; Virgili, M.; Orozco, M. Dissection of the Recognition Properties of p38 MAP Kinase. Determination of the Binding Mode of a New Pyridinyl−Heterocycle Inhibitor Family. *Journal of Medicinal Chemistry* **2007**, *50* (2), 283-293. DOI: 10.1021/jm061073h.
(73) Gelpí, J. L.; Kalko, S. G.; Barril, X.; Cirera, J.; de La Cruz, X.; Luque, F. J.; Orozco, M. Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins* **2001**, *45* (4), 428-437.
(74) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *Journal of Chemical Theory and Computation* **2007**, *3* (1), 156-169. DOI: 10.1021/ct600085e.
(75) Valdés-Tresanco, M. S.; Valdés-Tresanco, M. E.; Valiente, P. A.; Moreno, E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation* **2021**, *17* (10), 6281-6291. DOI: 10.1021/acs.jctc.1c00645.
(76) Sun, T.; Chen, Y.; Wen, Y.; Zhu, Z.; Li, M. PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. *Communications Biology* **2021**, *4* (1), 1311. DOI: 10.1038/s42003-021-02826-3.
(77) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports* **2016**, *6* (1), 29575. DOI: 10.1038/srep29575.
(78) Aldeghi, M.; Gapsys, V.; de Groot, B. L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Central Science* **2019**, *5* (8), 1468-1474. DOI: 10.1021/acscentsci.9b00590.
(79) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *Journal of Computational Chemistry* **2015**, *36* (5), 348-354. DOI: 10.1002/jcc.23804 (acccessed 2019/05/31).
(80) Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H. Calculation of Binding Free Energies. In *Molecular Modeling of Proteins*, Kukol, A. Ed.; Springer New York, 2015; pp 173-209.
(81) Seeliger, D.; de Groot, B. L. Protein thermostability calculations using alchemical free energy simulations. *Biophys J* **2010**, *98* (10), 2309-2316. DOI: 10.1016/j.bpj.2010.01.051.
(82) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem Sci* **2019**, *11* (4), 1140-1152. DOI: 10.1039/c9sc03754c.
(83) Tejedor, E.; Becerra, Y.; Alomar, G.; Queralt, A.; Badia, R. M.; Torres, J.; Cortes, T.; Labarta, J. PyCOMPSs: Parallel computational workflows in Python. *International Journal of High Performance Computing Applications* **2015**. DOI: 10.1177/1094342015594678.
(84) Hauser, K.; Negron, C.; Albanese, S. K.; Ray, S.; Steinbrecher, T.; Abel, R.; Chodera, J. D.; Wang, L. Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Commun Biol* **2018**, *1*, 70. DOI: 10.1038/s42003-018-0075-x.

(85) Galdadas, I.; Carlino, L.; Ward, R. A.; Hughes, S. J.; Haider, S.; Gervasio, F. L. Structural basis of the effect of activating mutations on the EGF receptor. *eLife* **2021**, *10*, e65824. DOI: 10.7554/eLife.65824.
(86) Sutto, L.; Gervasio, F. L. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proc Natl Acad Sci U S A* **2013**, *110* (26), 10616-10621. DOI: 10.1073/pnas.1221953110.
(87) Park, J.; McDonald, J. J.; Petter, R. C.; Houk, K. N. Molecular Dynamics Analysis of Binding of Kinase Inhibitors to WT EGFR and the T790M Mutant. *Journal of Chemical Theory and Computation* **2016**, *12* (4), 2066-2078. DOI: 10.1021/acs.jctc.5b01221.

TOC (7,16 cm × 18,47 cm)



TOC (3,28 cm × 8.46 cm)