



OPEN

The effect of protein mutations on drug binding suggests ensuing personalised drug selection

Shunzhou Wan^{1,5}, Deepak Kumar^{2,5}, Valentin Ilyin², Ussama Al Homsy³, Gulab Sher⁴, Alexander Knuth³ & Peter V. Coveney¹✉

The advent of personalised medicine promises a deeper understanding of mechanisms and therefore therapies. However, the connection between genomic sequences and clinical treatments is often unclear. We studied 50 breast cancer patients belonging to a population-cohort in the state of Qatar. From Sanger sequencing, we identified several new deleterious mutations in the estrogen receptor 1 gene (ESR1). The effect of these mutations on drug treatment in the protein target encoded by ESR1, namely the estrogen receptor, was achieved via rapid and accurate protein–ligand binding affinity interaction studies which were performed for the selected drugs and the natural ligand estrogen. Four nonsynonymous mutations in the ligand-binding domain were subjected to molecular dynamics simulation using absolute and relative binding free energy methods, leading to the ranking of the efficacy of six selected drugs for patients with the mutations. Our study shows that a personalised clinical decision system can be created by integrating an individual patient's genomic data at the molecular level within a computational pipeline which ranks the efficacy of binding of particular drugs to variant proteins.

Breast cancer is the most common cancer affecting women, and its mortality rate has increased significantly in the world during the past 25 years¹. Given the prevalence of breast cancer, it is pertinent that we devise high-throughput experimental and computational methods that provide a comprehensive and holistic understanding of the cause of cancer. In the post-genomic “one size does *not* fit all” era, personalised medicine is surely the way forward, considering the improved ability provided by the methodology to inform treatments that would work effectively for individual patients. Advances in genomic profiling of breast cancer have led to the identification of several key mutations in the disease^{2,3}. An in-depth understanding of the mechanisms of the disease requires not only a knowledge of the genome and its variants but the correct tools to fully interpret the knowledge. The pathways for the disease are routed via proteins, and it is their interactions that are amenable to treatment. This leads in turn to clinical decision support for personalised drug treatment. The lack of approved targeted treatments (other than mTOR inhibitors⁴ and anti-HER2 agents⁵), however, makes the genomic profiling of breast cancer less attractive compared with other tumours, such as lung cancer⁶.

An optimal selection of sequencing techniques is crucial to generate genomic libraries for specific patients, depending on sample size and the genomic targets. When interrogating a small region of DNA on a limited number of samples or genomic targets, Sanger sequencing is a good choice⁷. The estrogen receptor (ER) protein encoded by the ESR1 gene is expressed in about 70% of breast cancers⁸. ER also plays a vital role in classifying breast cancer subtypes and assigning therapeutic strategies; moreover, clinical research has established the central role of ER in the initiation and progression of breast cancers⁹. At least 62 ER mutations have been identified, of which most occur in the ligand-binding domain⁸. Several of the mutations are associated with ligand-independent ER activation or drug resistance. Experimental studies have revealed how some of the mutations affect the functions of ER, including changes of binding abilities for estradiol and drugs, abilities of dimerization, preferences of active and inactive states, and changes of interaction with cofactors and other proteins^{8,10}. Computational studies also show that ERs can be constitutively activated in their apo form by some mutations¹¹. Molecular Mechanics Poisson–Boltzmann Surface Area (MMPBSA) and Molecular Mechanics Generalized Born Surface Area (MMGBSA) approaches have also been used to study the binding free energy of ligands to

¹Department of Chemistry, Centre for Computational Science, University College London, London WC1H 0AJ, UK. ²Computational Biology, Carnegie Mellon University in Qatar (CMU-Q), Doha, Qatar. ³Hematology and Oncology Department, National Center for Cancer Care & Research, Hamad Medical Corporation, Doha, Qatar. ⁴Interim Translational Research Institute, Hamad Medical Corporation, Doha, Qatar. ⁵These authors contributed equally: Shunzhou Wan and Deepak Kumar. ✉email: p.v.coveney@ucl.ac.uk

the wild-type ER, although no overall correlation has hitherto been obtained between the calculations and the experimental results¹².

The significance of sequencing and sequenced data lies in the identification of biomarkers and aberrations in the genome profiles of breast cancer patients. Identification of mutations in the ESR1 gene through genome profiling dates back to Weis et al. in 1996, who addressed the effect of mutations on the conformational dynamics of the ER receptor¹³, and to Zhang et al. in 1997, who identified three missense mutations in a cohort of 30 tumours¹⁴. In our study we identified genetic aberrations in 50 breast cancer patients from a population cohort in the state of Qatar using Sanger sequencing targeted on ESR1, and performed ESMACS (enhanced sampling of molecular dynamics with approximation of continuum solvent)^{15,16} and TIES (thermodynamic integration with enhanced sampling)^{16,17} binding free energy studies to understand the effects of these mutations in a manner that could be used in the development of novel therapeutic strategies to inhibit these ER mutants and substantially improve treatment outcomes¹⁸. We have extensively validated the ESMACS and TIES approaches by applying them to a variety of proteins with diverse sets of ligands. The studies show that these ensemble-based approaches can generate precise and reliable free energy predictions, while TIES method is also accurate^{15–17,19–26}. We recently showed how such methods (i.e. ESMACS and TIES) can be used to assess functional and mechanistic impacts of mutations in the case of FGFR1 (fibroblast growth factor receptor 1) variants²⁷. In the longer term, a related approach could be used to design new drugs which are resistant to such mutations.

Materials and methods

Target gene sequencing for ESR1 gene was performed on the 50 tumour tissue samples collected from Qatari female patients with newly diagnosed estrogen receptor-positive breast cancer. The samples were preserved by formalin-fixed paraffin-embedded (FFPE) fixation. The study had been classified as “non-human subject research” by, and approvals granted from Institutional Review Board (IRB), Hospital Research Committee (HRC), Medical Research Center (MRC) and Hamad Medical Corporation (HMC) in Qatar. This ensured that we could deal with the anonymous tissue samples in accordance with relevant guidelines and regulations. Computational analysis was undertaken and missense SNP (single-nucleotide polymorphism) variants in the sequenced data were identified. The aim of the study was to identify possible ESR1 mutations within Qatari population, and to get an understanding of the drugs’ response to the potential mutations for future drug development and clinical treatment. The study was not conducted for the purpose of treatment of the patients from whom the samples were collected. Among the identified variants, four significant missense SNPs were used in further analysis to understand their effect on protein-drug interactions and protein activation using computer based molecular dynamics simulations. Because of the prospective nature of the modelling study, all of the four important SNPs are investigated, even though some of them have unclear chromatograms (see Table S3 in the Supplemental Material) and their statistical significance needs to be evaluated with large datasets.

Genome sequencing. Fifty breast cancer samples were collected from a population cohort of breast cancer patients in the state of Qatar at Hamad Medical Corporation (HMC) and were subjected to Sanger sequencing. The Sanger sequencing method was applied to the ten coding exons of the ESR1 gene in these samples to detect aberrant mutants. For sequencing, genomic DNA was isolated from formalin-fixed, paraffin-embedded tissue by Maxwell 16 FFPE Tissue LEV DNA Purification Kit (Promega). The quality and quantity of the DNA was checked by NanoDrop 2000c Spectrophotometer (Thermo Scientific) and agarose gel electrophoresis. Specific primers for the coding exons of the ESR1 gene (Transcript ID: ENST00000440973.5) were designed by Primer3web software, v4.0.0. The coding exons were amplified by PCR using Maxima Hot Start PCR Master Mix (Thermo Scientific) and purified by GeneJET PCR Purification kit (Thermo Scientific). Cycle sequencing was carried out using the BigDye Terminator v3.1 cycle sequencing kit. Sequencing reaction products were purified by the BigDye XTerminator purification Kit and analysed on an ABI 3500 Genetic Analyzer (Applied Biosystems). All procedures were carried out according to the manufacturers’ instructions. Finally, sequenced data was generated in ABI format files.

Thereafter, variant-calling computational analysis was performed on the sequenced data and missense variants were identified. These missense SNP variants in the sequenced data were called by SeqScape Software 3 (applied biosystems). A mutation report was generated for each patient. Chromatogram analysis was performed on the sequenced data to detect artefacts such as mis-called-nucleotides and aberrations. A list of SNPs (synonymous and nonsynonymous) was thus generated consisting of the patient number, mutation, and its novelty or known status based on variant databases: dbSNP²⁸, Ensemble²⁹, TCGA (<https://www.cancer.gov/tcga>), gnomAD³⁰, MOBCdb³¹, 1000 Genomes³², TOPMed³³, ExAC³⁴, COSMIC³⁵, HGMD³⁶ and ESP (<https://esp.gs.washington.edu/drupal/>). All new nsSNPs were analysed using StSNP³⁷.

Molecular dynamics based investigation of protein-drug interactions. Mutations obtained from sequencing analysis were subjected to a modelling and simulation study in order to understand the effect of these variants on the binding affinity of drugs to ER. To validate our computational approaches within the current molecular systems, a control study was also conducted, in which simulations were performed for three mutations—L387A, Y537S and D538G—of which experimental binding affinities were available^{10,38}.

We used ensemble-based ESMACS and TIES for the free energy calculations. Extensive studies have confirmed that the most effective and reliable computational route to reproducible predictions using MD simulation can be achieved using ensemble methods^{23,39–41}. A set of independent MD simulations are employed to obtain the required averages and associated uncertainties. The protocols for ESMACS and 5 replicas for TIES, with 4 ns production runs, were established in our previous studies^{15,17,20,23,39,42,43}, in which the number of replicas and the duration of the production runs were varied, and the results were compared between the ensemble runs

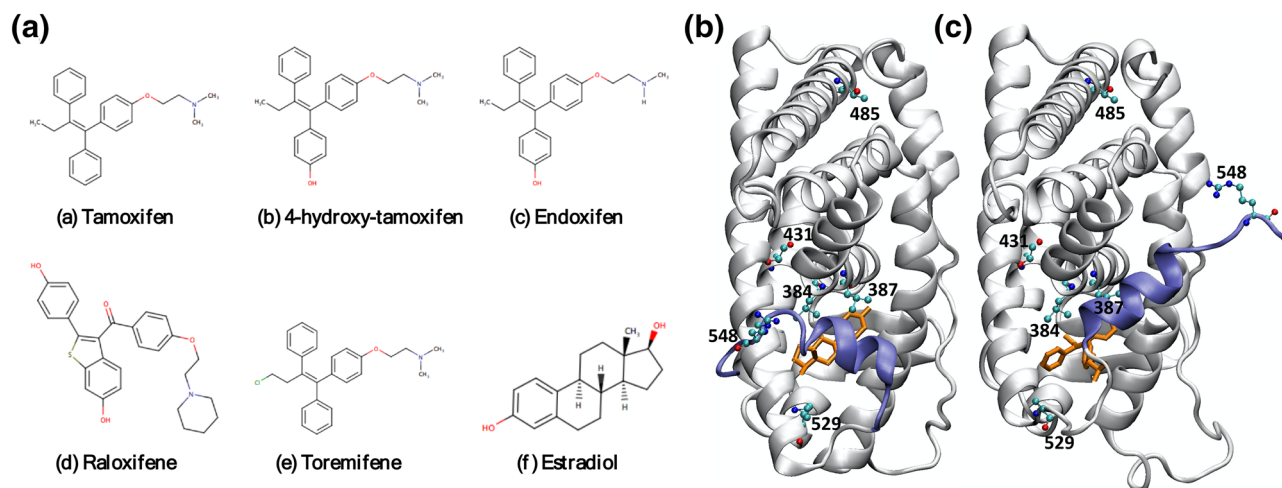


Figure 1. Chemical structures of the 6 ligands that have been investigated (a), and positions of the mutations identified from 50 breast cancer patients of Qatari nationals, in both the active (b) and inactive (c) conformations. The PDB code of the active conformation crystal structure is 1QKU, and the inactive conformation is 3ERT. The ligands presented in the crystal structures are represented as stick in orange, the protein is shown as cartoon in silver. The helix 12 (H12) is highlighted in blue, which shows different orientations in the active (b) and inactive (c) conformations.

and the “long time duration” single trajectory simulations. Our work demonstrates compellingly that the ensemble approach produced more precise and reproducible predictions than long simulations, even though the latter were several times longer in temporal duration than the entire ensemble simulation. The variations of the results from ensemble simulations are typically larger than those from single long simulations^{15,20,23,39}, indicating better conformational sampling achieved from the former. The following simulation study methodology was executed.

Molecular models. Binding affinities were obtained for 5 ER drugs or drug metabolites: toremifene (TOR), endoxifen (EDO), raloxifene (RAL), 4-hydroxy-tamoxifen (4-OHT) and tamoxifen (TMX), and the natural ligand estrogen (E2) for ER (Fig. 1a). The ligand-binding domain of the estrogen receptor is an α -helical bundle, of which several helices, particularly helix 12 (H12, see Fig. 1b,c), are known to be crucial for activity. At the active conformation, the H12 helix caps the ligand binding cavity (Fig. 1b) and its position is a prerequisite for coactivator recruitment to the activation function 2 (AF-2) cleft. In the inactive conformation, the H12 helix occupies the AF-2 cleft (Fig. 1c) preventing the coactivator to interact with the ER and to trigger transcription activity.

Two x-ray structures of the estrogen receptor, PDB codes 1QKU⁴⁴ and 3ERT⁴⁵, were used for this study, which represent the active and inactive forms of the protein, with the H12 helix at different positions (Fig. 1b,c). The ER structure of the former PDB model is complexed to 4-OHT, whereas the latter is bound to the native E2. 4-OHT and E2 bind to ER in different conformations. 4-OHT, an antagonist, displaces the usual position of the H12 helix so that the ER is found in an inactive conformation. E2, as the natural ligand for ER, fits in the binding pocket without sterically hindering the H12 helix, and thus the E2-ER complex exists in an active conformation (Fig. 1b). The complex structures for TOR, EDO and TMX were generated by replacing the 4-OHT inhibitor in 3ERT, after overlapping the common scaffold of the ligands. The coordinates of Ral in PDB 2QXS⁴⁶ were used to build the model of RAL after aligning the two PDB structures 2QXS and 3ERT. All corresponding crystallographic water molecules in 1QKU and 3ERT were retained.

ESMACS studies. Enhanced sampling of molecular dynamics with approximation of continuum solvent (ESMACS)^{15,16} studies employed an ensemble molecular dynamics approach which consists of 25 replica simulations. For each replica, the same initial coordinates were used for a given ligand-receptor complex, with different initial velocities randomly assigned to the atoms according to a Maxwell-Boltzmann distribution at 50 K. The systems were first heated over a period of 60 ps to 300 K, followed by 2 ns equilibration and 4 ns production runs for each replica. All simulations are performed in an isothermal-isobaric (constant temperature and constant pressure) ensemble using periodic boundary conditions. Free energy was evaluated approximately on the basis of the MMPBSA (molecular mechanics Poisson-Boltzmann surface area) method applied on a set of conformations from ensemble molecular dynamics simulations (see more details in the Supplemental Material).

TIES-PM studies. We have recently extended our TIES (thermodynamic integration with enhanced sampling) approach^{17,23} to study the free energy changes caused by protein mutations, a TIES variant we call TIES-PM¹⁹. We have established a standard protocol for TIES-PM, in which thirteen windows, consisting of the two endpoints representing the two physical states (WT and mutant ERs) and 11 intermediate states, are simulated for the alchemical process of protein mutation. The intermediate windows are mixtures of the two physical states

Residue number	Reference/mutant	Status	Mutant characterization	Comments ^a
384	L/V	Novel	real	Binding pocket
387	L/R	Novel	artefact	Binding pocket
431	T/A	Known	artefact	No direct interaction with the ligand
485	T/I	Novel	real	Far from binding site; may be important for domain-domain interaction or dimerization
529	K/N	Novel	artefact	At the C-terminal of helix H11, which links to the N-terminal of helix H12; may be important for the orientation of H12; not very far from the ligand (~7 Å)
548	R/P	Novel	artefact	At the C-terminal end of helix H12; may be important for the orientation of H12

Table 1. SNP missense variants obtained from Sanger sequencing study. ^aThe assumptions of their effects are based solely on the positions of the mutations in the static crystallographic structures (Fig. 1b,c). More studies on structures, dynamics and energetics will be required to confirm or refute these assumptions.

that consist of the appearing and disappearing parts of the residues (see Supplemental Material for more details). Simulations were performed for both ligand–protein complexes and apo-proteins. Five replicas were used for each window, from which the energy deviations and the statistical errors were calculated^{19,27}. The binding free energy differences were then calculated as the difference of the alchemical free energy changes in the apo-proteins and ligand bound complexes.

TIES-PM calculations involve an alchemical mutation between two amino acids. Four residue mutations identified in the current sequencing study were selected for the TIES-PM study: L384V, L387R, K529N and R548P. Although ESMACS and TIES (including TIES-PM) have been adequately validated for a variety number of protein systems, a control study is preferable here as no experimental data is available to support our predictions. We perform TIES-PM and ESMACS simulations for L387A, Y537S and D538G as an internal control. It should be noted that while L387A occurs inside the binding pocket (“local”), the other two mutations occur away from the binding pocket (“remote”). Our previous study has shown that alchemical methods, even with enhanced sampling approaches, may not be able to predict the binding free energy changes for such remote mutations¹⁹. Some of these mutations involve perturbing the net charge of the system, which requires additional calculations to take into account the resulting finite size electrostatic corrections to the free energy^{20,47}.

Simulations. The binding affinity calculator (BAC)⁴⁸ software tool was used to perform ESMACS and TIES studies. BAC constitutes a computational pipeline built from preparation and setup of the simulations, including parametrization of the compounds, solvation of the complexes, electrostatic neutralization of the systems by adding counterions and generation of configurations files for the simulations. The Amber package⁴⁹ was invoked for the setup of the systems and analyses of the results, and the MD package NAMD2.12⁵⁰ was used throughout the equilibration and production runs of all simulations. The AMBER ff99SBildn force field⁵¹ was used for the protein, and TIP3P was used for water molecules. Parameters for the ligands were produced using the general AMBER force field (GAFF)⁵² with Gaussian⁵³ calculations at the Hartree–Fock level with 6-31G** basis functions. The restrained electrostatic potential (RESP) module in the AMBER package⁴⁹ was used to calculate the partial atomic charges for the ligands. All of the ligands are electrostatically neutral except Ral which has a +1e net charge. All systems were solvated in orthorhombic water boxes with a minimum distance of 14 Å between box boundary and the ligand–protein complex. Standard protocols for ESMACS¹⁵ and TIES¹⁷ have been applied, in which simulations of multiple replicas were performed with identical initial conditions other than their initial velocities, which were drawn randomly from a Maxwell–Boltzmann distribution. Energy minimisation and 2 ns equilibration were conducted before 4 ns production runs were performed for each replica of the ESMACS and TIES-PM studies. Trajectories were recorded every 10 ps during the production runs for further analyses.

All simulations were run on the BlueWaters supercomputer at the National Center for Supercomputing Applications of the University of Illinois at Urbana–Champaign (<https://bluewaters.ncsa.illinois.edu>). Simulations of all replicas in an ensemble were executed concurrently, and completed in essentially the same amount of wall-clock time as that for one replica. For one single replica, a 2 ns equilibration and 4 ns production MD simulation took 15.7 h on 2 nodes (64 cores) of BlueWaters.

Results and discussion

Sequencing analysis. From our sequencing study, 22 mutations (Supplemental Material) were identified, of which six were nonsynonymous and in the ligand-binding domain, as shown in Table 1 and Fig. 1b,c. 7 of them were silent mutations, of which some were observed at relatively high frequencies (Table S2). Among these 22 mutations, 14 mutations were identified to be novel with no annotations available in nucleotide variants repositories. On the other hand, 8 mutations were found to be known with their respective annotations accessible in variant databases such as dbSNP. Corresponding frequencies of mutations in the studied 50 breast cancer samples were computed to understand their occurrence and cluster pattern across the analysed patient cohort (see the Supplemental Material). Such studies of detecting mutation occurrence patterns could be used for advance statistical analysis, where the identified mutations’ presence is not only studied in patient samples from a specific region but, also for its uniqueness and ubiquity in other assessed population cohorts. Furthermore, identification of unique and prevalent mutations in diverse ethnic populations will play a vital role toward the goal of precision medicine in pharmacology⁵⁴. Along these lines, data curation and mining were performed on breast cancer data available in public repositories to validate the novelty of the mutations identified with clear

Ligand	$\Delta\Delta G_{\text{ESMACS}}$		$\Delta\Delta G_{\text{TIES}}$		pdb
	L384V	L387R	L384V	L387R	
4-OHT	1.3 ± 1.1	1.0 ± 1.3	2.2 ± 0.4	5.1 ± 0.5	3ert
EDO	0.3 ± 1.1	0.9 ± 1.3	2.0 ± 0.4	4.8 ± 0.5	3ert
RAL	3.9 ± 1.9	3.6 ± 2.0	2.2 ± 0.4	6.1 ± 0.9	3ert
TMX	0.6 ± 1.0	3.4 ± 1.2	2.2 ± 0.4	4.7 ± 0.6	3ert
TOR	0.2 ± 1.0	4.3 ± 1.3	2.2 ± 0.5	4.6 ± 0.6	3ert
E2	1.7 ± 1.2	1.3 ± 1.3	2.2 ± 0.3	5.2 ± 1.8	1qku
Control E2	L387A	Y537S	D538G	pdb	
$\Delta\Delta G_{\text{ESMACS}}$	3.9 ± 1.3	1.0 ± 0.9	1.0 ± 0.9	1qku	
$\Delta\Delta G_{\text{TIES}}$	1.8 ± 0.2	-0.4 ± 0.5	-0.5 ± 0.6	1qku	
$\Delta\Delta G_{\text{exp}}$	0.5 ± 0.2 ¹⁰	1.1 ± 0.4 ³⁸	1.2 ± 0.4 ³⁸	-	

Table 2. Relative binding free energies $\Delta\Delta G = \Delta G_{\text{binding}}^{\text{mut}} - \Delta G_{\text{binding}}^{\text{WT}}$ for six ligands with two ER mutations—L384V and L387R—from ESMACS and TIES approaches. The calculations for the three mutations taken from the literature—L387A, Y537S and D538G—are presented as a control. The Poisson-Boltzmann (PB) free energy methods were used in the predictions of the ESMACS free energies, while alchemical approach was used for TIES. All energy values are in kcal/mol.

chromatograms (Table S3; Supplemental Material) in the Qatari breast cancer patient cohort studied. Among the curated databases, Ensemble is a comprehensive collection of variant information from multiple sources such as dbSNP, COSMIC, ESP and HGMD-PUBLIC. Moreover, the Ensemble database also provides evidences of the mutations' significance and validity from large-scale sequencing catalogues of human mutations and genotype data such as 1000 Genomes project, ExAC, TOPMed and gnomAD. From the analyses executed, it was deduced that the identified novel mutations hold their uniqueness among the cancer data present in the databases. A total of 226 synonymous and 373 non-synonymous mutations in ESR1 gene protein coding region were observed in the Ensemble repository, and their respective evidences were validated from 1000 Genomes project, ExAC, TOPMed and gnomAD. None of our identified novel mutations were reported in the analysed databases. Furthermore, novel mutations identified in Qatar cohort were not observed in the TCGA and MOCdb multi-omics breast cancer databases. A total of 17 samples with non-synonymous and 12 samples with synonymous somatic mutations in ESR1 gene were discerned in TCGA database—again, none of the novel mutations were observed in the studied TCGA samples. Additionally, ClinVar⁵⁵ database was mined to check the presence and clinical significance (if any) of the identified novel mutations; 13 mutations belonging to ESR1 gene with likely-pathogenic and pathogenic clinical significance status were noticed—no novel mutations from the analysed Qatar cohort were detected. Further, from the comprehensive Open Targets Consortium⁵⁶ consisting of evidence from genetics, genomics, transcriptomics and target-disease associations, we noticed that the identified novel mutations were not present in the Consortium database. Thus, from the available public data it can be concluded that the novel mutations identified in the Qatar breast cancer patient cohort have not been studied previously with any reported clinical significance. It should be noted that some mutational signatures are annotated as possible artefacts in Table 1 because of the chromatogram quality from the FFPE samples.

Two of the 22 mutations are at the binding site—L384V and L387R—while two located at or near helix 11 or 12—K529N and R548P—are important for the orientation of helix H12 (Fig. 1b,c). Interestingly, no mutations were found between amino acids 534–538, a region where most mutant residues are reported to cluster¹⁸.

These four mutations, L384V, L387R, K529N and R548P, are directly involved in ligand binding or protein activation, and were further investigated by our ESMACS and TIES approaches. The other two mutations, T431A and T485I, occur away from the ligand binding site or the helices H11/H12; these are not expected to affect the ligand binding or protein activation directly. They may modulate the protein stability and/or protein–protein interactions via induced allosteric conformational changes occurring over a wide range of space and time scales. The spatial and temporal scales are greater than standard atomistic molecular dynamics simulations can access¹⁹, and hence no further investigation was performed for these mutations using molecular dynamics modelling approaches.

Molecular dynamics study result. In the control study, the binding free energies of E2 were calculated for three mutations—L387A, Y537S and D538G, and compared with the experimental data^{10,38} (Table 2). The same binding assay was performed in the two publications^{10,38}, with different mutations. They both measured the dissociation constant of E2 with the wild-type ER, with results differing by more than 2 folds (equivalent to ~0.5 kcal/mol difference in the binding free energy). It highlighted the uncertainties of experimental measurement, and contributed to the differences between the calculations and experiments. For the local mutation L387A, the calculated binding free energy differences from ESMACS and TIES agree directionally with that from the experimental data; that is, both calculations and experiment show that the mutation weakens the binding of E2 to the protein. For the two remote mutations Y537S and D538G, the ESMACS approach correctly predicts the weakened binding. TIES approach, however, cannot predict such changes in binding affinities. As reasoned in our previous publication¹⁹, the effect of remote mutants affects the binding of a compound indirectly through an allosteric mechanism. TIES only samples local conformational changes which are not affected by

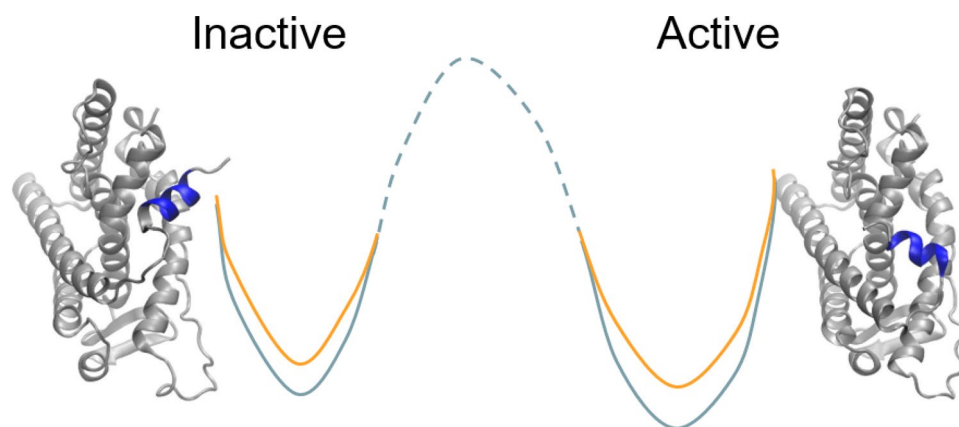


Figure 2. Conformational free energy changes of the active and inactive states due to mutation. The mutations change the relative energy differences between the two states, and hence shift the balance between them. The 2D energy surface illustrates an example of the energy changes at the two states, from wild-type (orange) to mutant (light blue), rendering the active state more favourable for the mutant protein than the wild-type. It should be noted that the 2D energy surface is an illustration as the free energy difference between the active and inactive conformations of the wild type, and the energy barrier between them (the dashed line) are unknown.

remote mutations¹⁹. Both ESMACS and TIES work well for local mutations; we therefore focus the free energy predictions on two mutations: L384V and L387R. It should be noted that while TIES approach is theoretically accurate, ESMACS invokes a few approximations in the energy estimations. ESMACS can generate rankings reasonably well for a set of compounds based on their binding affinities, but the differences of the affinities between pairs of compounds are not accurate.

Free energy calculations with ESMACS for mutations identified in Qatari population. The predicted binding affinities of L384V and L387R from ESMACS were compared with the wild-type results (Table 2). Other mutations (Table 1) were not included as they are positioned away from the binding pocket and do not have any direct interaction with the ligands. Some mutations, such as K529N and R548P, are located at a key position for the orientation of the helix H12 (Fig. 1b,c), and are expected to play a role in the active-inactive conformational changes. Their potential roles are investigated by the TIES approach (see the next section).

L384V and L387R induce resistance in all of the studied ligands, evidenced by the positive relative binding free energies for the mutated ERs compared with the wild-type ER with the corresponding ligands. L384V and L387R occur in the binding pocket and directly interact with the ligands. The L387R mutation, in particular, introduces not only steric bulk but a net electrostatic charge change. It induces significantly larger free energy changes for the ligands TMX and TOR than the mutation L384V does (Table 2). Large changes in the size of the residues and the charge distributions can confer resistance to and even completely block access to the ligands. To the best of our knowledge, there are no experimental data reported for the changes in binding affinity induced by these specific mutations. Other mutations, however, have been reported to weaken the binding of estradiol when they occur at the binding site¹⁰, including a mutation (L387A) occurring at the same position as L387R studied here.

Free energy calculations with TIES-PM for mutations identified in Qatari population. The binding affinities of two mutations L384V and L387R were investigated by the TIES-PM approach (Table 2). The L387R mutation involves a net charge change, and hence a finite-size effect needs to be taken into account^{20,47}.

Both mutations induce resistance, which is in line with the ESMACS predictions. The L384V mutation weakens the binding for all of the ligands by 2.0–2.2 kcal/mol universally. The L387R mutation has a higher impact on the ligand binding, reducing the binding affinities by 4.6–6.1 kcal/mol. In drug discovery, a rule of thumb to consider compounds for further development is to select those with dissociation constants (K_d) in the millimolar to micromolar range, usually with an equivalent binding affinity more negative than -6.5 kcal/mol⁵⁷. The large changes for the L387R mutation make the binding free energies all around or less negative than -6 kcal/mol for the ligands investigated here. This means that L387R mutation is likely to block the binding of all these ligands, including the native estradiol.

Conformation free energy changes with TIES-PM for mutations identified in Qatari population. The relative conformational free energy changes were investigated by the TIES-PM approach (Fig. S1b and Eq. S5 in Supplemental Material). Previous studies have shown that mutations can result in a change of activity for protein kinases. A gatekeeper mutation in fibroblast growth factor receptors (FGFRs), for example, has been shown to enhance the kinase activity using the later named TIES-PM approach²⁷. The estrogen receptor exists in at least two conformational states: active and inactive (Figs. 1b,c and 2). The receptor is likely to favour the inactive conformation at the physiological condition. Mutations may change the intrinsic equilibrium between the active and inactive states without ligand binding. Studies of large conformational changes are usually beyond the scope of standard

Mutation	Active		Inactive		$\Delta\Delta G$
	ΔG_{TIES}	ΔG_{FS}^*	ΔG_{TIES}	ΔG_{FS}^*	
L384V	2.7 ± 0.3	–	1.1 ± 0.4	–	1.6 ± 0.4
L387R	-29.6 ± 0.4	54.6	-37.8 ± 0.4	57.0	5.8 ± 0.5
K529N	30.8 ± 0.2	– 52.5	32.4 ± 0.7	– 55.3	1.2 ± 0.7
R548P	60.0 ± 0.7	– 52.5	56.3 ± 0.5	– 56.9	8.2 ± 0.8

Table 3. Relative conformational free energy changes $\Delta\Delta G = \Delta G_{TIES}^{act} - \Delta G_{TIES}^{inact}$ between the active and inactive states upon a mutation. $\Delta\Delta G > 0$ means that the free energy change in the active state is larger than that in the inactive state (Fig. 2). All energy values are in kcal/mol. *Finite size correction, related to the size of simulation box; the error associated is negligible.

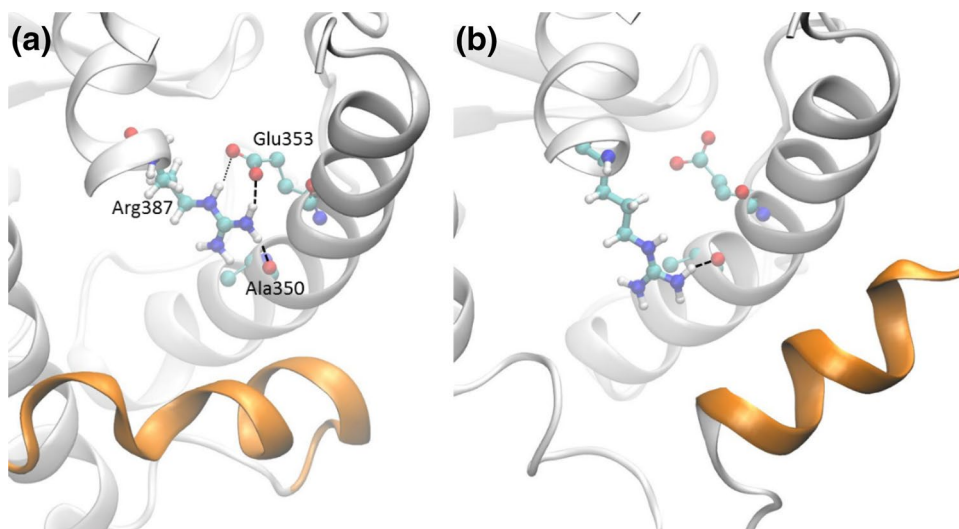


Figure 3. Formation of hydrogen bonds between the mutant residue 387 with other residues within the helix 3 (H3) at the active (a) and inactive (b) states. At the active state, Arg387 forms one stable hydrogen bond (bold dashed lines) with residues Ala350 and Glu353 each, and an additional one (light dotted line) with Glu353, which appears in ~ 65% of the entire simulations. At the inactive state, the side chain of Arg387 only forms one stable hydrogen bond with Ala350 (side chain of Glu353 forms hydrogen bonds with Arg394 instead, with 126% frequency of occurrence; the frequency is 15% in the active state). The helices H4, H8 and H9 are removed from both figures for reasons of clarity.

all-atom molecular dynamics simulation¹⁹; while “accelerated” MD simulations can provide a free energy profile between the two states²⁷, they come with large uncertainties owing to the nature of the approximations used.

The TIES-PM approach can deliver accurate and precise predictions, and is used here to investigate the relative binding free energy changes in the two states caused by protein mutations (Fig. 2). The apo forms of the protein are simulated at both the active and inactive states. For each state, TIES-PM is performed to achemically transfer the protein from wild-type to mutant form (Fig. S1b). The alchemical free energy changes are then used to calculate the conformational free energy difference $\Delta\Delta G$ (Eq. S5). $\Delta\Delta G$ is a physical property which is used here to quantify the changes of the preference for the two states. The calculations show that the L384V and K529N mutations confer a moderate change on the relative stability of these two protein states, rendering the active state slightly more favourable, or less unfavourable, for the mutant protein than the wild-type (Table 3). By contrast, the L387R and R548P mutations have a large impact on the preference of the two states, making the active state significantly more favourable, or less unfavourable, for the mutant proteins than for the wild-type. It should be noted that the finite size electrostatic corrections contribute importantly to these calculations and improve the predicted free energy changes significantly.

Structural base for the preference of active state. Our free energy results showed that, thermodynamically, all of the 4 mutations prefer the active state over the wild type. For the wild-type protein, the residue Leu387 participates in hydrogen bonding (see more details in the Supplemental Material) only via its main chain atoms to form the α -helix structure. It enjoys a similar pattern of hydrogen bonding, with 79% and 69% frequencies of occurrence in the active and inactive states, respectively. The substitution of Leu387 with a positively charged, polar residue Arg387 creates more hydrogen bonds via its side chain atoms (Fig. 3). In the active state, the side chain atoms form hydrogen bonds with residues in helix 3, with a frequency of 222% (2.22 hydrogen bonds on average, see Fig. 3a). In the inactive state, the H12 helix packs with helices H3 and H5, and slightly changes the

orientation and conformation of the latter. The residue Glu358 on H3, which maintains stable hydrogen bond with Arg387 in the active state, forms hydrogen bond with Arg394 instead in the inactive conformation. As a result, the side chain of Arg387 only maintain one hydrogen bond, with a frequency of 103%, with residues in helix 3 (see Fig. 3b). The more stabilising hydrogen bonds in the active state shift the balance between the active and inactive forms, making the L387R variant thermodynamically preferable to the active state.

For the other mutations, the reasons for the free energy changes are more subtle, and cannot be ascribed to any single, dominant contribution. It is likely, however, that for the L384V variant, a less bulky substituent reduces the steric hindrance of the H12 helix when the protein is in the active conformation (Fig. 1b). For the mutations K529N and R548P, which are both located at the surface of the protein and involve net charge changes, it is likely that the stability of the protein is affected mainly by electrostatic interactions and solvation effects. The stability of the protein is probably attribute to the conformations and energetically of the side chains, as no significant changes are observed in the residue-wise root mean square fluctuations for the main chain atoms. For these mutations, there may not be one single indicator that explains why either the active or inactive state is favoured.

Conclusions

Our study, performed on 50 breast cancer patients in a Qatari population cohort, furnishes a holistic understanding of the effect of deleterious mutations on the effectiveness of prevalent breast cancer drugs available today. Moreover, although the present study is based on a small set of 50 breast cancer patients, it demonstrates the power of patient-specific medical approaches in treating breast cancer as it reveals the presence of uncommon mutations among patients within one local and small geographical region. The sequencing study identified several mutations among breast cancer patients in Qatar. Some of these mutations are of considerable interest, and have not been previously reported in the public repositories of cancer data. In the future, in tandem with the validation of the identified novel mutations in the Qatari population cohort from publicly available consortiums, we would like to collect more samples, both within Qatar and worldwide, to perform computational analysis and determine whether these novel mutations are specific to the Qatari population and to investigate their more general importance.

Based on this genomic analysis, we then performed a rigorous and in depth molecular modelling study of the estrogen receptor with sequential variations obtained from the gene sequencing study in this project. The molecular modelling approaches were applied to the newly identified mutations in the ligand-binding domain of the receptor. The predicted binding free energies provide a clear explanation for the effects of these mutations. The mutations at the binding site, L384V and L387R, induce resistance to the drugs studied here; the mutations L387R and R548P play an important role in the activation of the estrogen receptor. This methodology may in the future be employed as the basis for a clinical decision support tool for patient specific drug treatment: the combination of rapid genome sequencing and binding affinity calculations offers a powerful and reliable way to provide patient specific treatment regimens. Along similar lines, these approaches may also be used to design new drugs which inhibit the development of resistance in the target proteins.

Received: 2 January 2021; Accepted: 9 June 2021

Published online: 29 June 2021

References

1. Azamjah, N., Soltan-Zadeh, Y. & Zayeri, F. Global trend of breast cancer mortality rate: A 25-year study. *Asian Pac. J. Cancer Prev.* **20**, 2015–2020. <https://doi.org/10.31557/APJCP.2019.20.7.2015> (2019).
2. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479. <https://doi.org/10.1038/ncomms11479> (2016).
3. Giltane, J. M. *et al.* Genomic profiling of ER+ breast cancers after short-term estrogen suppression reveals alterations associated with endocrine resistance. *Sci. Transl. Med.* **9**, eaai7993. <https://doi.org/10.1126/scitranslmed.aai7993> (2017).
4. Xie, J., Wang, X. & Proud, C. G. mTOR inhibitors in cancer therapy. *F1000Res* **5**, 2078. <https://doi.org/10.12688/f1000research.9207.1> (2016).
5. Tai, W., Mahato, R. & Cheng, K. The role of HER2 in cancer therapy and targeted drug delivery. *J. Control Release* **146**, 264–275. <https://doi.org/10.1016/j.jconrel.2010.04.009> (2010).
6. Pinto, J. A. *et al.* Precision medicine for locally advanced breast cancer: Frontiers and challenges in Latin America. *Ecancermedicalscience* **13**, 896. <https://doi.org/10.3332/ecancer.2019.896> (2019).
7. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8. <https://doi.org/10.1016/j.ygeno.2015.11.003> (2016).
8. Dustin, D., Gu, G. & Fuqua, S. A. W. ESR1 mutations in breast cancer. *Cancer* **125**, 3714–3728. <https://doi.org/10.1002/cncr.32345> (2019).
9. Huang, B., Warner, M. & Gustafsson, J. A. Estrogen receptors in breast carcinogenesis and endocrine therapy. *Mol. Cell Endocrinol.* **418**(Pt 3), 240–244. <https://doi.org/10.1016/j.mce.2014.11.015> (2015).
10. Metivier, R. *et al.* A dynamic structural model for estrogen receptor-alpha activation by ligands, emphasizing the role of interactions between distant A and E domains. *Mol. Cell* **10**, 1019–1032. [https://doi.org/10.1016/s1097-2765\(02\)00746-3](https://doi.org/10.1016/s1097-2765(02)00746-3) (2002).
11. Pavlin, M. *et al.* A Computational assay of estrogen receptor alpha antagonists reveals the key common structural traits of drugs effectively fighting refractory breast cancers. *Sci. Rep.* **8**, 649. <https://doi.org/10.1038/s41598-017-17364-4> (2018).
12. Liu, J. Y. & Mooney, S. D. Characterization of ligand type of estrogen receptor by MD simulation and mm-PBSA free energy analysis. *Int. J. Biochem. Mol. Biol.* **2**, 190–198 (2011).
13. Weis, K. E., Ekena, K., Thomas, J. A., Lazennec, G. & Katzenellenbogen, B. S. Constitutively active human estrogen receptors containing amino acid substitutions for tyrosine 537 in the receptor protein. *Mol. Endocrinol.* **10**, 1388–1398. <https://doi.org/10.1210/mend.10.11.8923465> (1996).
14. Zhang, Q. X., Borg, A., Wolf, D. M., Oesterreich, S. & Fuqua, S. A. An estrogen receptor mutant with strong hormone-independent activity from a metastatic breast cancer. *Cancer Res.* **57**, 1244–1249 (1997).

15. Wan, S., Knapp, B., Wright, D. W., Deane, C. M. & Coveney, P. V. Rapid, precise, and reproducible prediction of peptide-MHC binding affinities from molecular dynamics that correlate well with experiment. *J. Chem. Theory Comput.* **11**, 3346–3356. <https://doi.org/10.1021/acs.jctc.5b00179> (2015).
16. Wan, S., Bhati, A. P., Zasada, S. J. & Coveney, P. V. Rapid, accurate, precise and reproducible ligand-protein binding free energy prediction. *Interface Focus* **10**, 20200007. <https://doi.org/10.1098/rsfs.2020.0007> (2020).
17. Bhati, A. P., Wan, S., Wright, D. W. & Coveney, P. V. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.* **13**, 210–222. <https://doi.org/10.1021/acs.jctc.6b00979> (2017).
18. Reinert, T., Goncalves, R. & Bines, J. Implications of ESR1 mutations in hormone receptor-positive breast cancer. *Curr. Treat. Opt. Oncol.* **19**, 24. <https://doi.org/10.1007/s11864-018-0542-0> (2018).
19. Bhati, A. P., Wan, S. & Coveney, P. V. Ensemble-based replica exchange alchemical free energy methods: The effect of protein mutations on inhibitor binding. *J. Chem. Theory Comput.* **15**, 1265–1277. <https://doi.org/10.1021/acs.jctc.8b01118> (2019).
20. Bhati, A. P., Wan, S., Hu, Y., Sherborne, B. & Coveney, P. V. Uncertainty quantification in alchemical free energy methods. *J. Chem. Theory Comput.* **14**, 2867–2880. <https://doi.org/10.1021/acs.jctc.7b01143> (2018).
21. Wan, S. *et al.* Evaluation and characterization of Trk kinase inhibitors for the treatment of pain: Reliable binding affinity predictions from theory and computation. *J. Chem. Inf. Model* **57**, 897–909. <https://doi.org/10.1021/acs.jcim.6b00780> (2017).
22. Wan, S. *et al.* Rapid and reliable binding affinity prediction of bromodomain inhibitors: A computational study. *J. Chem. Theory Comput.* **13**, 784–795. <https://doi.org/10.1021/acs.jctc.6b00794> (2017).
23. Wan, S., Tresadern, G., Pérez-Benito, L., Vlijmen, H. & Coveney, P. V. Accuracy and precision of alchemical relative free energy predictions with and without replica-exchange. *Adv. Theory Simul.* **3**, 1900195. <https://doi.org/10.1002/adts.201900195> (2019).
24. Wright, D. W. *et al.* Application of ESMACS binding free energy protocols to diverse datasets: Bromodomain-containing protein 4. *Sci. Rep.* **9**, 6017. <https://doi.org/10.1038/s41598-019-41758-1> (2019).
25. Wright, D. W. *et al.* Application of the ESMACS binding free energy protocol to a multi-binding site lactate dehydrogenase A ligand dataset. *Adv. Theory Simul.* **3**, 1900194. <https://doi.org/10.1002/adts.201900194> (2019).
26. Wan, S. *et al.* Hit-to-lead and lead optimization binding free energy calculations for G protein-coupled receptors. *Interface Focus*, **10**, 20190128. <https://doi.org/10.1098/rsfs.2019.0128> (2020).
27. Bunney, T. D. *et al.* The effect of mutations on drug sensitivity and kinase activity of fibroblast growth factor receptors: A combined experimental and theoretical study. *EBioMedicine* **2**, 194–204. <https://doi.org/10.1016/j.ebiom.2015.02.009> (2015).
28. Sherry, S. T. *et al.* dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311. <https://doi.org/10.1093/nar/29.1.308> (2001).
29. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761. <https://doi.org/10.1093/nar/gkx1098> (2018).
30. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* <https://doi.org/10.1101/531210> (2019).
31. Xie, B. *et al.* MOBCdb: A comprehensive database integrating multi-omics data on breast cancer for precision medicine. *Breast Cancer Res. Treat.* **169**, 625–632. <https://doi.org/10.1007/s10549-018-4708-z> (2018).
32. Genomes Project, C *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74. <https://doi.org/10.1038/nature15393> (2015).
33. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* <https://doi.org/10.1101/563866> (2019).
34. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291. <https://doi.org/10.1038/nature19057> (2016).
35. Tate, J. G. *et al.* COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947. <https://doi.org/10.1093/nar/gky1015> (2019).
36. Stenson, P. D. *et al.* Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581. <https://doi.org/10.1002/humu.10212> (2003).
37. Uzun, A., Leslin, C. M., Abyzov, A. & Ilyin, V. Structure SNP (StSNP): A web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Res.* **35**, W384–392. <https://doi.org/10.1093/nar/gkm232> (2007).
38. Zhao, Y. *et al.* Structurally novel antiestrogens elicit differential responses from constitutively active mutant estrogen receptors in breast cancer cells and tumors. *Cancer Res.* **77**, 5602–5613. <https://doi.org/10.1158/0008-5472.CAN-17-1265> (2017).
39. Coveney, P. V. & Wan, S. On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys. Chem. Chem. Phys.* **18**, 30236–30240. <https://doi.org/10.1039/c6cp02349e> (2016).
40. Knapp, B., Ospina, L. & Deane, C. M. Avoiding false positive conclusions in molecular simulation: The importance of replicas. *J. Chem. Theory Comput.* **14**, 6127–6138. <https://doi.org/10.1021/acs.jctc.8b00391> (2018).
41. Wan, S., Sinclair, R. C. & Coveney, P. V. Uncertainty quantification in classical molecular dynamics. *Phil. Trans. R. Soc. A*, **379**, 20200082. <https://doi.org/10.1098/rsta.2020.0082> (2021).
42. Bieniek, M. K., Bhati, A. P., Wan, S. & Coveney, P. V. TIES 20: Relative binding free energy with a flexible superimposition algorithm and partial ring morphing. *J. Chem. Theory Comput.* **17**, 1250–1265. <https://doi.org/10.1021/acs.jctc.0c01179> (2021).
43. Wright, D. W., Hall, B. A., Kenway, O. A., Jha, S. & Coveney, P. V. Computing clinically relevant binding free energies of HIV-1 protease inhibitors. *J. Chem. Theory Comput.* **10**, 1228–1241. <https://doi.org/10.1021/ct4007037> (2014).
44. Gangloff, M. *et al.* Crystal structure of a mutant hERalpha ligand-binding domain reveals key structural features for the mechanism of partial agonism. *J. Biol. Chem.* **276**, 15059–15065. <https://doi.org/10.1074/jbc.M009870200> (2001).
45. Shiau, A. K. *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927–937 (1998).
46. Bruning, J. B. *et al.* Coupling of receptor conformation and ligand orientation determine graded activity. *Nat. Chem. Biol.* **6**, 837–843. <https://doi.org/10.1038/nchembio.451> (2010).
47. Olsson, M. A., Garcia-Sosa, A. T. & Ryde, U. Binding affinities of the farnesoid X receptor in the D3R Grand Challenge 2 estimated by free-energy perturbation and docking. *J. Comput. Aided Mol. Des.* **32**, 211–224. <https://doi.org/10.1007/s10822-017-0056-z> (2018).
48. Sadiq, S. K. *et al.* Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 proteases. *J. Chem. Inf. Model* **48**, 1909–1919. <https://doi.org/10.1021/ci8000937> (2008).
49. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688. <https://doi.org/10.1002/jcc.20290> (2005).
50. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802. <https://doi.org/10.1002/jcc.20289> (2005).
51. Wang, B., Li, L., Hurley, T. D. & Meroueh, S. O. Molecular recognition in a diverse set of protein-ligand interactions studied with molecular dynamics simulations and end-point free energy calculations. *J. Chem. Inf. Model* **53**, 2659–2670. <https://doi.org/10.1021/ci400312v> (2013).
52. 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.* **25**, 1157–1174. <https://doi.org/10.1002/jcc.20035> (2004).
53. Frisch, M. J. *et al.* *Gaussian 03* (Gaussian, Inc., Wallingford, CT, 2004).
54. Cascorbi, I. Significance of pharmacogenomics in precision medicine. *Clin. Pharmacol. Ther.* **103**, 732–735. <https://doi.org/10.1002/cpt.1052> (2018).

55. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985. <https://doi.org/10.1093/nar/gkt1113> (2014).
56. Carvalho-Silva, D. *et al.* Open Targets Platform: New developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065. <https://doi.org/10.1093/nar/gky1133> (2019).
57. Chodera, J. D. & Mobley, D. L. Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**, 121–142. <https://doi.org/10.1146/annurev-biophys-083012-130318> (2013).

Acknowledgements

The authors would like to acknowledge (i) Qatar National Research Fund (Grant No. 7-1083-1-191), (ii) the UK Medical Research Council for funding the Medical Bioinformatics project (MR/L016311/1), (iii) EU H2020 projects ComPat (<http://www.compat-project.eu/>, Grant No. 671564), CompBioMed and CompBioMed2 (<http://www.compbioimed.eu>, Grant Nos 675451 and 823712), (iv) NSF Award (<https://www.nsf.gov/pubs/2017/nsf17542/nsf17542.htm>, Award No. NSF 1713749) and (v) special funding to PVC from the UCL Provost, and (vi) Dr. Zafar Nawaz, Vinod Kumar Gupta, Dr. Imaad bin Mujeeb and Dr. Cicy Mary Jacob from National Centre for Cancer Care and Research, Hamad Medical Corporation for providing breast cancer samples for the study. We made use of the BlueWaters supercomputer at the National Center for Supercomputing Applications of the University of Illinois at Urbana-Champaign (<https://bluewaters.ncsa.illinois.edu>), access to which was made available through the aforementioned NSF award, and the Titan supercomputer at the Oak Ridge National Laboratory, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Author contributions

S.W.: Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, review & editing. D.K.: Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, review & editing. V.I.: Funding acquisition, Conceptualization, Supervision, Writing—review & editing. U.A.H.: Funding acquisition, Conceptualization, Supervision, Writing—review & editing. G.S.: Data curation. A.K.: Clinical and medical advice, Writing—review & editing. P.V.C.: Funding acquisition, Conceptualization, Methodology, Supervision, Writing—original draft, review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-92785-w>.

Correspondence and requests for materials should be addressed to P.V.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021