Towards Personalised Drug Ranking in Clinical Decision Support

By David W. Wright, Shunzhou Wan†, S. Kashif Sadiq, Stefan J. Zasada and Peter V. Coveney‡

Centre for Computational Science, Department of Chemistry,University College London, London, WC1H 0AJ, U.K.

Many infectious diseases as well as cancers are strongly influenced by molecular level processes. In several cases, the advent of rapid genetic sequencing, already available in the case of HIV, means that patient-specific treatment based on genetic data becomes conceivable. Targeted therapies use drugs to interfere with specific biomacromolecules involved in disease development. Given the complexity of emergent mutations in such biomacromolecules and in the disease itself, clinicians need to resort to decision support software for patient-specific treatment. Incorporating model based molecular level information into such decision support systems offers the potential to substantially enhance personalised drug treatment by providing first principles based ranking of drug efficacy on a specific patient. Patient specific molecular models of targeted macromolecules are constructed and molecular dynamics simulations are used to rank drug binding affinities. Here we present results from clinically relevant protein variants that arise from two distinct pathologies: HIV and lung carcinoma. Our findings demonstrate the potential for molecular simulations to achieve an accurate ranking of drug binding affinities on clinically relevant time scales and represent the first steps towards the eventual goal of providing data derived from patient specific simulation to enhance clinical decision support systems. The approach gives rapid, robust, and accurate computational results and is dependent on an automated workflow for building, simulating and analysing models distributed over petascale computing resources which are comprised of tens to hundreds of thousands of compute cores.

Keywords: Molecular Dynamics, Patient Specific Medicine, Clinical Decision Support, HIV-1, Protease, Lopinavir, Cancer, EGFR, Gefitinib

1. Introduction

Clinical decision support systems (CDSS) have been widely promoted as a means of processing available information and retrieving protocols for diagnosis, staging personalised treatment and follow-up with the overall aim of improving patient outcomes[12]. The general framework is based on the statistical profiling of patients in order to find matching genotypes and phenotypes. Personalised healthcare is recommended to new patients according to these profiles. Most of the efforts made so far focus on diagnosis based on clinical features and treatment based on

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

‡ Corresponding author: p.v.coveney@ucl.ac.uk

clinical practice. With the explosion in genomic and proteomic data, an increasing number of features can be included in CDSS. Indeed, CDSS are developing into environments that provide tools to integrate clinical and genomic features, assess the quality of recommendations, and evaluate the efficiency of the computer aided diagnosis and treatment.

The many possible choices of drugs in some diseases, and the promising progress made in pharmaceutical development for others, invite the prospect of incorporating drug ranking into CDSS in order to predict drug sensitivity and resistance at the genotypic level. In the HIV case, several existing CDSS are in widespread use today, such as Stanford HIVdb (`hivdb.stanford.edu`), ANRS (`www.anrs.fr`) and RegaDB (`www.rega.kuleuven.be/cev/regadb/`, the contents of which are dependent on the gathering of information from the published literature and expert opinion. Studies assessing these systems indicate that while performance is of a generally high standard†, with little difference in accuracy between CDSS for sequences frequently observed in patients, over 30% of sequences exhibit at least minor discordances in the level of drug susceptibility assigned by these different systems [9, 31]. The idea that computational modelling could be used to enhance or complement such systems has been widely discussed [22]. In response the EU FP6 funded ViroLab project (`http://www.virolab.org`) [41] developed a prototype clinical decision support system which differs from those currently available by incorporating a Popperian approach for personalised drug ranking (Figure 1). An automated workflow supplements the pre-existing 'Baconian' decision support systems with Popperian predictive modelling and drug ranking protocols based on molecular simulations [5]. We emphasize, however, that at present simulation is not employed in patient care and traditional, purely 'Baconian' CDSS remain the current approach for resistance interpretation.

Personalised drug ranking studies require access to both appropriate patient data and an integrated IT infrastructure linking such data to high performance computing (HPC) resources through fast networks. To this end, we are collaborating with clinicians who provide access to patient specific genomic data on HIV/AIDS and lung cancer cases. We have developed a highly automated molecular simulation based free energy calculation workflow tool, the Binding Affinity Calculator (BAC) [34], to perform drug ranking with optimal efficiency. We have integrated our Application Hosting Environment (AHE) [50] with BAC, so that applications can be launched automatically on numerous HPC resources on geographically distributed grids and federations of grids. The AHE is a lightweight hosting environment for running applications on grid resources: it provides a high level of abstraction for simulations and analyses of molecular level drug-protein interactions, choreographing the vast number of steps, including data transfers and production molecular dynamics that demand access to tens of thousands of cores on petascale compute resources, which in totality constitute our workflow. The input to this workflow is a specified drug and target protein combination along with the mutations present in the patient derived sequence relative to a defined wildtype. In the HIV domain the use of viral genetic data from individual patients is already routine and we envisage our simulations using pre-existing systems and protocols to acquire the

† Performance of CDSS was assessed by looking at the success in reducing viral load to undetectable levels 12, 24 and 48 weeks after a change in treatment determined after CDSS consultation.

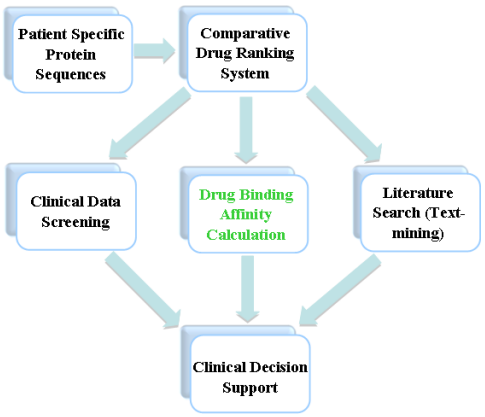


Figure 1: The architecture of a clinical decision support system incorporating molecular simulation. CDSS consists of three main problem solving components - patient clinical database search, drug binding affinity from molecular simulations, and literature mining - that share a common genetic knowledge of the individual patient. Our approach (green) uses MD simulations to estimate the binding affinity of various drugs with their targeted protein, and ranks drug efficacy for patients who have a specific protein variant. The decision support kernel integrates both the simulation and text mining information into integrated decision support for drug ranking.

necessary sequence information [41]. CDSS are less advanced for cancer drug targets; in response to the need to facilitate coordinated access to such patient genetic data the Individualised MEdiciNe Simulation Environment (IMENSE) [51] has been developed. This system was developed as part of the EU FP7 VPH project ContraCancrum (`http://www.contracancrum.eu/`) which aims to develop a composite multilevel platform for simulating malignant tumour development, along with tumour and normal tissue response to therapeutic modalities and treatment schedules in order to optimise the disease treatment procedure in the patient's individualized context. The simulations of proteins and drugs relevant to the treatment of lung cancer presented in this paper represent the smallest biological length and time scales involved in this process.

2. Drug Resistance Rankings From Molecular Simulation

The incorporation of predictive models into CDSS requires three main problems to be overcome: identification of a metric correlated to clinical response which can be computed from simulation, generation of sequence specific models, and the entire workflow to be turned around on a clinically relevant timescale.

The first consideration is to select a metric to assess the resistance level of a particular genetic sequence. The clinical impact of mutations is determined by a number of factors, including the strength of drug binding but also of changes in enzymatic efficacy and interactions with other host or disease factors. In the case of HIV-1 there is evidence that genotype to phenotype mapping correlates strongly with clinically observed outcomes [11]. The assays upon which these conclusions are founded concentrate upon drug binding alone and do not consider other potential

confounding factors. While the experiments upon which these studies are based are too time consuming and expensive to be applied routinely in the clinical context, the results suggest that measurements of the strength of drug binding would be a useful, predictive metric to obtain from simulations.

In order to quantify the strength of drug binding it is necessary to consider the underlying physics of binding. The binding of reactants at constant temperature and pressure is driven by the minimisation of the thermodynamic potential known as the Gibbs free energy, G. The strength of protein ligand binding is characterised by the change in this potential, ΔG, (also known as the binding free energy) which is given by:

$$\Delta G = \Delta H - T\Delta S \tag{2.1}$$

at thermodynamic temperature T, where ΔH is the change in enthalpy and ΔS the change in entropy upon binding. The more negative the ΔG value, the more tightly a drug binds to its target. Any attempt to evaluate the relative strength of drug binding equates to an estimate of the changes in ΔG. In this paper we have use the term 'binding affinity' as synonymous with the binding free energy, ΔG; however it is also widely used to refer to the equilibrium association constant for drug and protein, K_a. The two quantities are related via the van't Hoff equation:

$$\Delta G = -RT \ln K_a \tag{2.2}$$

where R is the gas constant and T the thermodynamic temperature. A change in binding free energy of 1.4 kcal mol^{-1} corresponds to a 10 fold change in K_a; changes of this magnitude result in significantly reduced inhibitor efficacy.

The ultimate cause of differences in the binding affinity residues in changes to the structure and chemical character of target proteins induced by alterations to their sequence. In order to describe these features we require molecular models of the protein-drug interactions of interest. The basis of any such modelling must be experimentally generated structures (usually derived from x-ray crystallography). In general, however, there are many more possible sequences of interest than available crystal structures; therefore mutations must be inserted *in silico*, a process known as homology modelling [32]. In the case of HIV, a variety of studies have been conducted which attempt to predict resistance levels from models of protein structure [6, 17, 40]. Such studies, based on molecular docking techniques, have had some success but fail to account for several factors that play important roles in determining binding strength in many situations [20, 26]. One such factor, protein flexibility and dynamics, can be accounted for using molecular dynamics (MD) simulations, in which atoms are characterised by their mass, partial charge and bonding characteristics, while Newtonian mechanics is used to evolve the system and allow the sampling of relevant protein conformations. Recently, much evidence has built up suggesting that the application of MD may help to improve the accuracy and reproducibility of binding affinity estimates [16, 26, 42], and it is this technique upon which the present article focuses. It is important to recognise, however, that MD is a computationally expensive technique: to produce results on a clinically relevant timescale requires the exploitation of petascale supercomputing resources [33, 36].

In order to be considered as viable candidates for incorporation in CDSS it is vital that binding affinity calculations from MD simulations are not only validated using comparisons to experimental findings but are also reproducible. Whilst many earlier MD studies claim agreement with experimental binding affinity values, it is often hard to determine whether this is representative of the simulation protocol and free energy calculation method or merely fortuitous. In this paper we report simulations of the HIV-1 protease and the human epidermal growth factor receptor (EGFR), implicated in the development of lung carcinoma, which address both of these issues. In addition we present an example of how such simulations could be used to evaluate resistance levels for a HIV-1 protease sequence in which existing CDSS produce ambiguous results. These findings provide some of the groundwork necessary to fully validate this approach in readiness for its use in clinical settings.

3. Binding Affinity Calculator

In order to be included as part of a CDSS the results of MD simulations must not only be reliable and available in a timely manner but the clinician should not have to be aware of the complicated workflow used to produce them. This requirement prompted the development of the Binding Affinity Calculator (BAC) [34]. Originally developed to study drugs targetted at the HIV-1 protease, BAC has now been extended for investigation of drugs targetted against the HIV-1 reverse transcriptase and EGFR. Here we describe the simulation and analysis protocol utilised by BAC and the infrastructure and middleware that it exploits. We then discuss two examples of its use within the HIV-1 protease and EGFR systems.

(*a*) *Evaluation of the Binding Affinty*

Many approaches are available for calculating binding affinities from MD simulations ranging from the theoretically exact, such as thermodynamic integration (TI), to the largely empirical, such as the linear interaction energy (LIE) method (excellent reviews of the subject are available by Gilson and Zhou [14], and Steinbrecher and Labahn [42]). The computational requirements of these methods tend to increase considerably as more physical detail is included in the model. The CDSS context means that simulation results must be turned around on a timescale of only a few days. In order to fulfill this requirement we employ the approximate Molecular Mechanics Poisson Boltzmann solvent accessible Surface Area (MMPBSA) methodology [21, 25] which provides a compromise between rapidity and accuracy of calculation. This method possesses several limitations for computing absolute binding free energies. It does not implicitly account for free energy differences that arise due to conformational changes upon binding, possible variations in key protonation states, and changes due to explicit water-mediated binding between protein and ligand, all of which can provide significant contributions to the binding free energy [26, 46]. Despite these limitations, our previous work demonstrates that changes in binding affinity of less than 1 kcal mol^{-1} between HIV-1 protease mutants can be distinguished [43]. Closer agreement with experimental binding affinity values can be achieved by incorporating a normal mode (NMODE) [1] estimation of the entropic component of the binding free energy.

Both MMPBSA and NMODE computations are applied to configuration snapshots generated over the course of MD simulations. The absolute free energy difference of binding, ΔG_{theor}, calculated using this methodology is given by:

$$\Delta G_{theor} = \langle \Delta H_{theor} \rangle_M - \langle T\Delta S_{theor} \rangle_N \tag{3.1}$$

Here, $\langle \Delta H_{theor} \rangle_M$ denotes the average of the enthalpically dominated MMPBSA calculation over M snapshots, while $\langle \Delta S_{theor} \rangle_N$ denotes the average change in configurational entropy resulting from NMODE calculations across N snapshots. Enthalpies and configurational entropies were calculated at a frequency of 100 and 5 snapshots/ns respectively.

The enthalpic value of each snapshot is given by:

$$\Delta H_{theor} = \Delta H_{vdW}^{MM} + \Delta H_{ele}^{MM} + \Delta H_{pol}^{sol} + \Delta H_{nonpol}^{sol} \tag{3.2}$$

where ΔH_{vdW}^{MM} and ΔH_{ele}^{MM} are the van der Waals and electrostatic contributions to the molecular mechanics free energy difference, respectively, and ΔH_{pol}^{sol} and ΔH_{nonpol}^{sol} are the polar and nonpolar solvation terms, respectively.

The molecular mechanics free energy differences (ΔH_{vdW}^{MM} and ΔH_{ele}^{MM}) were calculated using the SANDER module in AMBER 9 [4], with no cutoff for the non-bonded energies. The AMBER PBSA module was used to solve the linearized Poisson-Boltzmann equation to evaluate the electrostatic free energy of solvation ΔH_{pol}^{sol}. The nonpolar solvation free energy ΔH_{nonpol}^{sol} was calculated from the solvent accessible surface area (SASA) using the MSMS program [37]. Normal mode calculations were performed in the AMBER NMODE module. Full details of the parameters used by BAC are give in Sadiq et al. [34].

(*b*) *Model Preparation*

Collections of the parameters used to describe atoms within MD simulations are known as forcefields, of which there are several well established examples available to describe the amino acid constituents of proteins [15]. Once a structure of the sequence has been generated, mass, charge and bonding parameters are assigned to each atom in the structure. Forcefield parameters for drug compounds are not, in general, included in standard forcefields and must be added seperately. The standard AMBER force field for bioorganic systems (ff03) [7] provided the protein parameters. Drug coordinates were extracted from the appropriate crystal structures and missing hydrogens incorporated using the PRODRG tool [39]. Gaussian 03 [10] was used to perform geometric optimisation of all inhibitors at the HartreeFock level with 6-31G** basis functions. The restrained electrostatic potential (RESP) procedure, which is part of the AMBER9 package [4], was used to calculate the partial atomic charges. The force field parameters for the inhibitors were completely described by the general AMBER force field (GAFF) [45].

(*c*) *Simulation Protocol*

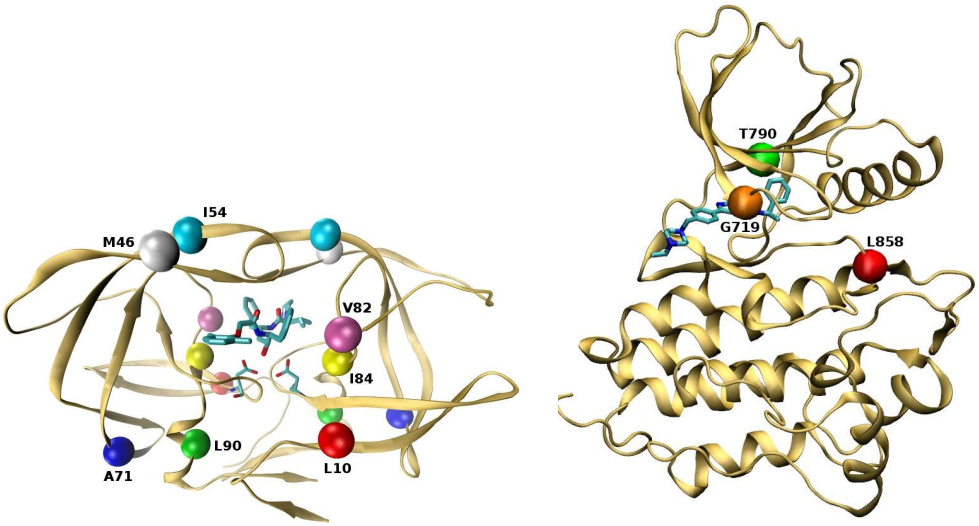
All simulations presented here were performed in the molecular dynamics package NAMD2 [30] using the protocol incorporated into the BAC software (based on

that originally employed by Perryman et al. [29]) which has previously been successfully used to calculate binding free energies for a number of inhibitors bound to various HIV protease sequences [35, 43]. Each protein sequence is solvated in a cuboid box of TIP3P water molecules [18], with a minimum buffering distance of 14 Å in all three orthogonal dimensions. The system is then minimised with all protein and ligand heavy atoms constrained to their positions in the initial structure. Each system is heated from 50 to 300 K over 50 ps and then maintained at a temperature of 300 K. Once the system reaches the correct temperature in all subsequent simulation steps the pressure is maintained at 1 bar. This results in the system sampling an isothermal isobaric (NPT) ensemble. The simulation proceeds for 200 ps before a mutation relaxation protocol is enacted. The relaxation protocol consists of the sequential release of constraints on each mutated residue (together with any residue within 5 Å) for 50 ps (constraints are maintained on the rest of the protein structure). This allows the residues to reorientate into more favourable conformations if necessary. After the 50 ps relaxation period the restraints are reapplied to each region. The final equilibration stage is the gradual reduction of the restraining force on the complex from 4 to 0 kcal $\mathrm{mol}^{-1}\mathrm{Å}^{-2}$ during a 350 ps period. Following this, the systems are allowed to evolve freely (a more detailed description of the equilibration protocol is given in the Supplementary Information). The entire equilibration stage is designed to take 2 ns for all systems, meaning that this final stage varies in length according to the number of mutations that require relaxation in the previous stages. After the equilibration is complete, structures are output for analysis every picosecond. Each output snapshot is post processed using MMPBSA, meaning that a hundred sets of coordinates are analysed for each nanosecond of simulation. The more computationally expensive NMODE analysis is performed on every 20 snapshots, producing five entropy estimates per nanosecond of simulation. A detailed description of the setup and simulation protocol is provided in Sadiq et al. [34].

A major challenge in the computational calculation of binding affinities is to obtain sufficient sampling of the energy landscape to produce converged results. Recent studies in our group [35] and by others [13] have indicated that using ensembles of short simulations with subtly different initial conditions reduces the wallclock time taken to meet this requirement compared to computing single long trajectories. As a result of this observation, all free energy values reported in the following studies were obtained from ensembles of 50 replica simulations, generating 4ns of production simulation, varying from one another only in the velocities initially assigned to the atoms in the simulation.

(*d*) *Computational Infrastructure and Middleware*

One essential requirement of a CDSS is that the validity of drug ranking relies not only on the correctness of the results, but on its timeliness. To support patient-specific medical care, the employed computers must be capable of running very large scale simulations within the time frames required in a clinical context [36]. The ensemble approach, dividing each calculation into a set of small replica simulations, lends itself well to the utilisation of distributed resources such as those available on the US Teragrid (`www.teragrid.org`), UK National Grid Service (`www.ngs.ac.uk`) and EU DEISA (`www.deisa.eu`) grids. The execution of a large number of replicas



(a) (b)

Figure 2: Three dimensional structures of the proteins simulated in the studies described in this paper. The backbone of both proteins is shown in ribbon representation with the locations of the mutations under investigation indicated by coloured balls and bound drugs using stick representation. (a) HIV-1 protease, bound to the inhibitor lopinavir (the catalytic dyad is shown in stick representation). The locations of the mutations found in the multi drug resistant (MDR) mutants (described in Table 1) used for the benchmark simulations and residue A71 are labelled. Protease is a homodimer and the location of each mutation is given the same colour on both monomers. (b) EGFR bound to the inhibitor AEE788 with the locations of G719, T790, and L858 highlighted.

in parallel also provides a significant improvement in terms of the turnaround time compared to running a single longer simulation. The approach is greatly facilitated by the current generation of petascale supercomputers which offer many tens of thousands of cores (planned future development of exoscale systems with many millions of cores will make the technique even more facile). The drawback of the ensemble approach is the need to manage the data for each replica individually. The BAC transparently implements data transfer and access to remote computational resources. BAC has recently been extended to take advantage of the Pilot-Job functionality provided in SAGA (Simple API for Grid Applications) to further enhance the efficiency of resource utilisation [23].

4. HIV and Lung Carcinoma

We have studied two systems where there is potential for molecular simulations to enhance future CDSS. The first is HIV-1 protease, for which traditional CDSS are well established; the second is epidermal growth factor receptor (EGFR), a drug target in lung cancer for which CDSS tools are only now being developed. Structures of both target proteins are shown in Figure 2.

(*a*) *HIV-1 Infection*

As part of the EU ViroLab project we helped to create a prototype CDSS which provides a common environment for the integration of simulations ranging from the molecular to the population levels alongside traditional rules based systems [2, 41] (a demo of the system is available at the ViroLab portal: `https://portal.virolab.org`). To show the potential of integrating diverse systems such as traditional drug ranking systems, literature mining, patient data and predictive simulations into a single interface, a so-called Virtual Patient Experiment (VPE) was designed. The aim of the VPE is to take a patient sequence for which the ViroLab comparative drug ranking system (cDRS) provided discordant results for one of the available protease inhibitors and to use the other tools within the system to produce insight that could help a clinician facing a decision on how to treat this virtual patient. This scenario highlights a situation in which simulation could be particularly useful in enhancing resistance assessments; namely, when the much more rapid statistical CDSS approach (results from which may be available in a matter of minutes or seconds) fails to produce an unambiguous rating, perhaps due to the rarity of the mutations (or the particular pattern of mutations) present.

The ViroLab cDRS allows the user to simultaneously obtain drug resistance rankings (susceptible, intermediate or resistant) for an input sequence or set of mutations from three well established drug ranking systems: Stanford HIVdb, ANRS and RegaDB.† The ViroLab and EuResist (`www.euresist.org`) databases were queried to find patient sequences that met these criteria, resulting in the choice of a sequence containing the mutations L10I, I13V, K14R, I15V, K20T, L63P, A71IV, V77I, L90M, I93L in combination with the drug lopinavir. This sequence was deemed to be susceptible by HIVdb but displayed intermediate resistance according to ANRS and Rega. In instances such as this, the Virtual Laboratory (VL) provides a tool which allows a clinician or researcher to investigate the cause of the discordance by inspecting the rules used to determine the ranking by each system. The only mutations within this set that influenced the ranking were L10I, A71IV and L90M. At position 71, where two mutations were detected in the patient, we chose to focus our investigation on the isoleucene substitution as it is rarer and so more likely to be less well represented in the data built into the existing CDSS. Here we briefly describe results used to assess the accuracy and reproducibility of binding affinity estimates produced by our free energy calculation protocol (full details are available in a recently published paper by Sadiq et al. [35]). We also present our contribution to the VPE as a vignette illustrating one way in which BAC could be used in conjunction with existing CDSS.

In order to establish the efficacy of our simulation and analysis protocol we conducted a study in which we attempted to replicate the experimental results achieved by Ohtaka et al. [27] on the HIV-1 protease inhibitor lopinavir (LPV) bound to a series of multi drug resistant (MDR) protease mutants. Alongside the HXB2 wild-type five variants containing subsets of a of a group of six mutations, namely L10I, M46I, I54V, V82A, I84V and L90M, with varying degrees of resistance were considered. The subsets of mutations have been labelled with two letter codes, shown in Table 1; this nomenclature will be used for the remainder of this report. HIV-1

† The following versions of the drug ranking systems' rule sets were used in determining the sequence to be investigated in the VPE: HIVdb 5.1.2, ANRS 17 and Rega 8.0.1.

Code	Description	Mutations
WT	Wildtype	HXB2
HM	MDR hexa-mutant	L10I, M46I, I54V, V82A, I84V, L90M
QM	MDR quatro-mutant	M46I, I54V, V82A, I84V
AS	Active site mutant	V82A, I84V
FL	Flap mutant	M46I, I54V
DM	Dimer interface mutant	L10I, L90M

Table 1: Two letter codes and sequence composition for the protease sequences of the multi drug resistant (MDR) mutants used to evaluate a suitable simulation protocol.

protease is a homodimer and hence a single dimeric mutation corresponds to two amino acid mutations, one at each identical position along the monomer. As can be seen in Figure 3, we achieve excellent agreement between our computed results and the experimental values, obtaining a correlation coefficient of 0.98 for both the results including and excluding the normal mode estimate of the entropic contribution (ΔG_{theor} and ΔH_{theor} respectively). Whilst the overall correlation is not affected by inclusion of the entropic contribution it is necessary to reproduce the experimental rank order of the variants. The minimal difference made to the correlation reflects the fact that the change in entropy, ΔS, is similar for LPV binding to all sequences, hence the inter system difference, $\Delta\Delta S$ is generally negligible.

In order to assess the reproducibility of our results we performed further separate ensembles of the WT and HM systems. Results for both were within 0.70 kcal mol^{-1} for ΔH_{theor} and 0.82 kcal mol^{-1} for ΔG_{theor}, respectively. These values should be contrasted with the range of values obtain for the constituent replicas within the ensembles where the largest differences between runs were 17.58 kcal mol^{-1} for ΔH_{theor} and 29.10 kcal mol^{-1} for ΔG_{theor}. The contrast in these values emphasizes the impact of even tiny changes in initial conditions on the final results of simulations and suggests that many reports of agreement between simulation derived free energy calculations and experiment may be fortuitous.

In addition to the ranking of the variants bound to LPV we also correctly calculated the binding affinity difference, of approximately 3 kcal mol^{-1}, between the binding of LPV and the less well optimised inhibitor saquinavir to the wildtype enzyme.

The binding affinities we compute exclude contributions from changes in conformation upon drug binding, alteration of the catalytic dyad protonation state and binding of a conserved water molecule. However, the correlation with experiment suggests that these are not substantially altered by the resistance associated mutations. Nonetheless, the discrepancy with experiment meant that we used the calculated values for the susceptible wildtype, WT, and highly resistant, HM, systems as benchmarks when making predictions of resistance in other sequences.

The ViroLab VL is designed to facilitate basic research as well as to provide a platform to enhance CDSS. In this context, BAC is a software tool which can be used not only to rank mutant sequences in terms of resistance but also to provide some level of insight into the way which different mutations within the sequence combine, in the hope that this will shed light on the origin of inconsistencies in data derived from other sources. With this in mind it was decided that, rather

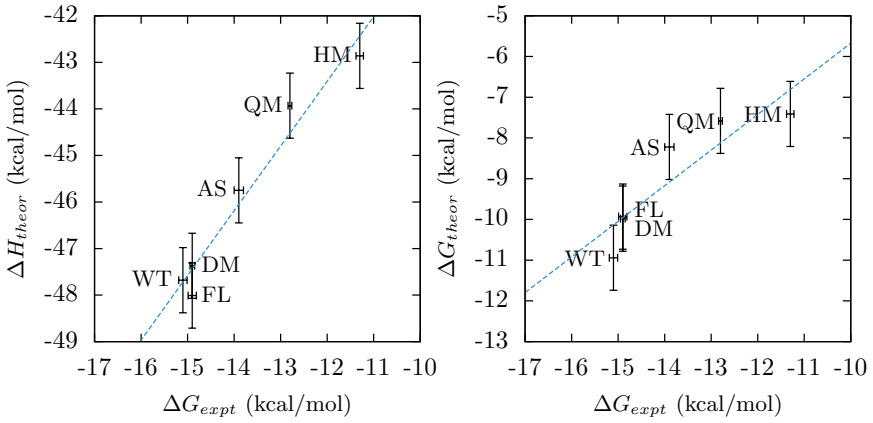
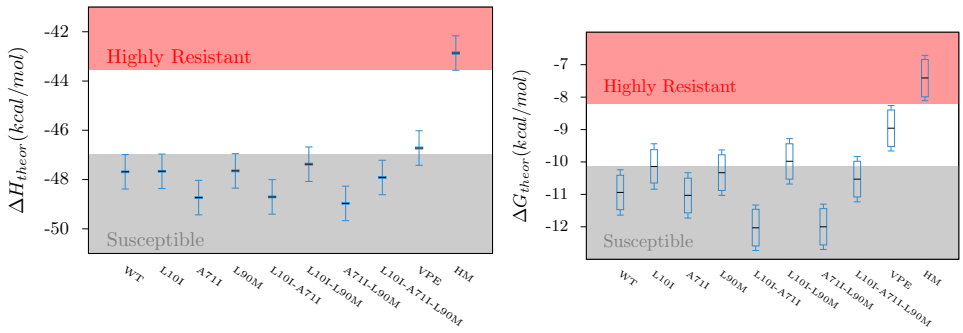


Figure 3: Comparison of average computed binding affinities with those derived experimentally by Ohtaka *et al.* [27]: a) ΔH_{theor} is the enthalpically dominated binding affinity from MMPBSA calculation alone; b) ΔG_{theor} is the absolute binding affinity with the entropic component calculated from normal mode analysis. The blue line represents a linear regression performed on each data set, both of which exhibit a correlation coefficient of 0.98 between the computed and experimental values. The error bars of 0.7 kcal mol^{-1} for ΔH_{theor} and 0.8 kcal mol^{-1} for ΔG_{theor} were derived from a reproducibility study of the WT and HM systems as reported in Sadiq et al. [35].

than simply simulating the L10I, A71I, L90M mutation set, we would simulate *all* possible combinations of the three constituent point mutations along with the full patient sequence (which we label as VPE).

Figure 4 shows binding affinity results from BAC analysis of these sequences bound to LPV, with the seven mutant sequences under investigation compared to the established WT and HM benchmarks for susceptibility and high level resistance. A number of observations can be made from these graphs. Firstly, there are notable differences between the entropy and absolute binding affinity results for the triple mutant L10I-A71I-L90M and the complete VPE sequence. Using both measures, the triple mutant would be regarded as susceptible whereas the VPE sequence is distinctly ranked as resistant using ΔG_{theor} (the difference between compared to wildtype, $\Delta\Delta G_{theor}$, is 1.98 kcal mol^{-1}), while the change in enthalpy also suggests some level of resistance ($\Delta\Delta H_{theor}$ is 0.96 kcal mol^{-1}). None of the other mutational combinations are evaluated as causing any resistance, with the possible exception of the L10I-L90M sequence. The L10I-L90M system has a $\Delta\Delta G_{theor}$ of 0.96kcal mol^{-1} which indicates some level of resistance but this is close to the limit of resolution of our method, and the $\Delta\Delta H_{theor}$ of 0.30kcal mol^{-1} would be classified as susceptible. The A71I, L10I-A71I and A71I-L90M have $\Delta\Delta H_{theor}$ values of -1.05, -1.02 and -1.29 kcal mol^{-1} respectively, all of which are greater than the reproducibility variation in observed in the WT and HM systems. This increase in the strength of binding is conserved for the L10I-A71I and A71I-L90M systems when the entropic contribution is included in the results (they have $\Delta\Delta G_{theor}$ values of -1.09 and -1.06 kcal mol^{-1} respectively), but not for the A71I single



(a) (b)

Figure 4: A comparison of the computed binding affinities for all combinations of the mutations L10I, A71I and L90M and the full VPE sequence with the known susceptible WT sequence and known resistant HM sequence. (a) shows the binding enthalphy, ΔH_{theor}, alone and (b) the absolute free energy difference, ΔG_{theor}. The black lines show the mean, the candle sticks the standard error and the whiskers the error based on the WT and HM reproducibility for each system. The grey and red shaded regions show the range of values deemed susceptible and resistant respectively, defined using the WT and HM benchmark values.

mutant (which is almost indistinguishable from WT with a $\Delta\Delta G_{theor}$ of -0.09 kcal mol^{-1}). This suggests that at least the double mutants containing A71I may be hyper-susceptible to LPV. The phenomenon of hyper-susceptibility to LPV has been observed experimentally in a range of sequences although the clinical impact remains unknown [24, 47].

These observations suggest a possible explanation for the discordance found using the cDRS. The effects on drug binding of the mutations at positions 10, 71 and 90 appear to be highly dependent on the background of other mutations present in the sequence. If the mutations are rare then the subtle intragenic epistatic effects which cause this phenomenon are unlikely to present themselves frequently enough to be picked up as statistically meaningful in the databases used to establish the CDSS rules. The non-additive nature of the interactions is likely to be a further confounding factor, as linear regression based techniques are often used in establishing the rule sets [22].

(*b*) *Lung Cancer*

Lung cancer is the leading cancer-related cause of death worldwide [28]. It is usually treated by a combination of chemotherapy alongside other treatments like radiation therapy and surgery. Targeted therapy [38] is a new approach to cancer treatment, which is expected to be more effective and less harmful than traditional chemotherapies. The approach uses drugs to interfere with a specific molecular target, usually a protein with a critical role in tumour growth. Epidermal growth factor receptor (EGFR) is such a drug target in lung cancer, and possibly in other forms of cancer, because it is frequently overexpressed and/or overactive in cancerous cells [52]. EGFR is a membrane-spanning cell surface protein. Its intracellular tyrosine kinase domain is a preferred target for small compounds tyrosine kinase inhibitor (TKIs) to inhibit the kinase activity and suppress its function. Clinical

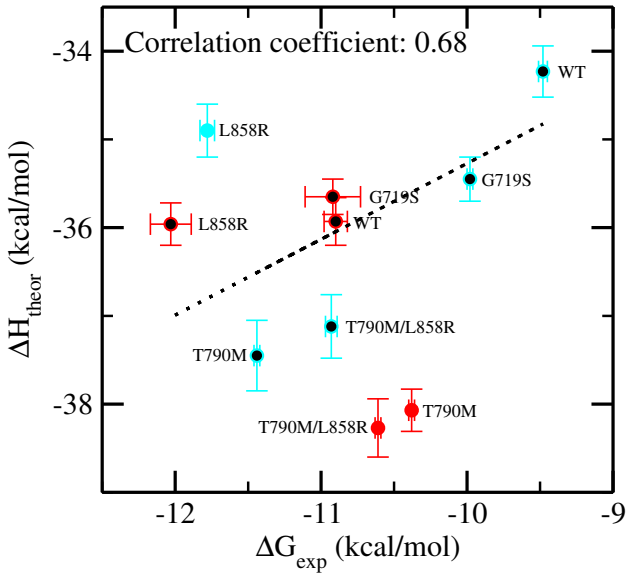


Figure 5: Comparison of calculated binding enthalpies (ΔH_{theor}) with experimental binding free energies (ΔG_{exp}) for inhibitors AEE788 (red) and Gefitinib (blue) complexed with EGFRs. Seven points (those in black) are used for linear fitting. The error bars are shown as standard errors of the mean from ensemble simulations and standard deviations from experiments [48, 49]. Gefitinib-L858R, AEE788-T790M and AEE788-T790M/L858R are excluded from linear fitting as they may be outliers (see text for discussion).

studies manifest a strong correlation between the presence of mutations and patient response to TKIs. However, genotypic assaying is not routinely performed for cancer patients, in contradistinction with the HIV case discussed in the previous section.

Reversible TKIs compete with ATP binding to the kinase, and hence prevent the phosphorylation of EGFR. There are three clinically approved TKIs for EGFR: Gefitinib, Erlotinib and Lapatinib. A number of other TKIs are currently in various stages of clinical trials. In clinical use on nonsmall cell lung cancer patients, Gefitinib is found to be effective in individuals with specific somatic mutations. EGFR tyrosine kinase domain mutations are usually clustered around the ATP-binding pocket. They can distort the ATP-binding cleft and change the relative binding affinity of the kinase domain for inhibitors and ATP. A greater understanding of how inhibitors interact with their target protein could lead to better optimised choice of drug treatments and/or selections of patient subgroups. In this section, we investigate the binding properties of two inhibitors Gefitinib and AEE788 with wild-type and four mutant EGFRs, namely G719S, L858R, T790M and T790M/L858R. The L858R is the most frequently found point mutation in sequences of individuals that respond to Gefitinib treatment, while T790M is found in Gefitinib-resistant nonsmall cell lung cancer patients.

In Figure 5, the calculated binding energies ΔH_{theor} are compared with the experimental data [48, 49] for Gefitinib and AEE788. When combining calculated binding energies of two inhibitors and comparing with experimental data [48, 49],

a reasonable correlation is obtained by excluding three data points (Figure 5). The Gefitinib-L858R appears to be an outlier, since it lies between Gefitinib-WT and Gefitinib-G719S in another experiment [8]. The experimental binding free energies for AEE788-T790M and AEE788-T790M/L858R are also dubious as they are not in line with a recent publication [19]. The physico-chemical properties of the inhibitors are the determinants for specific binding. Hydrogen bonds are one of the key interactions between an inhibitor and its targetted protein. It is reasonable to assume that AEE788 has greater binding affinities (more negative binding energies) to all forms of EGFR than Gefitinib does, as AEE788 forms two hydrogen bonds with EGFR while Gefitinib has only one. Our calculations and the experimental data [19] both confirm this, and hence raise questions as to the quality of the published binding affinities for AEE788-T790M and AEE788-T790M/L858R [49]. The simulations confirm that our ensemble method is able to produce consistent and reproducible results even when different starting structures are used. In contrast, considerable variances exist between experimental measurements made under different conditions. More complete discussions are presented elsewhere for the method and its application in lung cancer [44].

A cross-drug correlation makes it possible to identify subgroups of patients who have a specific EGFR variant and are most likely to respond well to a particular drug treatment, and to choose a personalised drug therapy that maximizes treatment efficacy for an individual. The effects of genetic changes on the overall protein structure are usually small; however, they are critical for drug binding, and can render previously susceptible proteins untargettable. Hence, targets need to be defined more specifically and precisely, with consideration given not only to the choice of molecule but also the particular genetic variants present in individual patients. Our theoretical predictions could be better evaluated in future given access to large scale genetic and clinical data from programmes of this kind. These results indicate that it would be beneficial for cancer patients to have genotypic assays performed in a similar way to that which is routine for HIV/AIDS patients today.

One of the aims of the EU ContraCancrum project is the creation of a data warehouse collating data from both experimental and *in silico* sources which may be used to inform future CDSS. Unlike the HIV case, genotypic testing is not currently standard for patients presenting with cancer; consequently a much more limited range of data is available. Encouragingly, however, the U.K. National Health Service (NHS) has recently announced plans to deploy broad genetic testing for people with various forms of cancer, including lung carcinoma, and to implement personalised medicine based on individuals' genetic information [3]. The program will enrol up to 12,000 patients in its first phase, many more than any other current clinical trials for cancer treatment.

5. Conclusions

Personalised drug ranking is an important component of clinical decision support that predicts drug sensitivity and resistance for individual patients. Many challenges remain before free energies from molecular dynamics simulations can be routinely used as part of CDSS but we have demonstrated the potential of such an approach to accurately rank drug binding affinities on clinically relevant timescales (2-3 days). In particular we have demonstrated the importance of the use of en-

semble simulations in order to obtain converged and reproducible free energies in contrast to single simulations which produce results that can be highly variable. Importantly, our methodology does not contain any adjustable parameters which have to be fitted meaning that the approach can be used to make predictions in a wide variety of systems. We can then verify the validity of our predictions against the available experimental data. Obtaining the well converged free energy values reported here has involved the production and analysis of many terabytes of data and was only possible by using large scale compute resources. It is to be hoped that the deployment of the next generation of exoscale machines will make this level of sampling possible on a routine basis. These developments pave the way for possible, future, use of simulations and free energy calculations in clinical decision support tools that match treatments to individual patients' genetic profiles. With the advent of even faster and cheaper genetic sequencing, such an approach should serve to further enhance outcomes based on individualized treatment, and to shape future clinical decision support systems that will provide more reliable healthcare.

6. Acknowledgements

This work has been supported in part by the EU FP7 ContraCancrum (ICT-2007.5.30), the EU FP6 ViroLab (IST-027446) and the EU FP7 CHAIN (HEALTH-2007-2.3.2-7) projects. Access to US TeraGrid was made available by the National Science Foundation under NRAC grant MCA04N014, and by the DEISA Consortium (co-funded by the EU, FP7 project 222919) to supercomputers in Europe via the Virtual Physiological Human Virtual Community managed by the VPH Network of Excellence (IST-223920: `www.vph-noe.eu`). We are grateful to EPSRC and CoMPLEX for funding the Ph.D. studentships of D.W.W. We also wish to acknowledge the UK NGS for providing access to their resources and their support for our work.

References

[1] B. R. Brooks, D. Janezic, and M. Karplus. Harmonic analysis of large systems. I. Methodology. *J. Comput. Chem.*, 16:1522–1542, 1995. doi: 10.1002/jcc.540161209.

[2] M. Bubak, T. Gubala, M. Malawski, B. Balis, W. Funika, T. Bartynski, E. Ciepiela, D. Harezlak, M. Kasztelnik, J. Kocot, D. Krol, P. Nowakowski, M. Pelczar, J. Wach, M. Assel, and A. Tirado-Ramos. Virtual laboratory for development and execution of biomedical collaborative applications. *IEEE Symposium on Computer-Based Medical Systems*, pages 373–378, 2008. ISSN 1063-7125. doi: 10.1109/CBMS.2008.47.

[3] E. Callaway. Cancer-gene testing ramps up. *Nature*, 467:766–767, 2010. doi: 10.1038/467766a.

[4] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–1688, 2005. doi: 10.1002/jcc.20290.

[5] P. V. Coveney and P. W. Fowler. Modelling biological complexity: a physical scientist's perspective. *J. R. Soc. Interface*, 2(4):267–280, 2005. doi: 10.1098/rsif.2005.0045.

[6] S. Draghici and R. B. Potter. Predicting HIV drug resistance with neural networks. *Bioinformatics*, 19(1):98–107, Jan 2003. doi: 10.1093/bioinformatics/19.1.98.

[7] Y. Duan, C. Wu., S. Chowdhury., M. C. Lee, G. Xiong, W. Zhang., R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003. ISSN 0192-8651. doi: 10.1002/jcc.10349. URL `http://dx.doi.org/10.1002/jcc.10349`.

[8] M. A. Fabian, W. H. Biggs, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko, A. G. Lai, J. Lelias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, and D. J. Lockhart. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.*, 23:329–336, 2005. doi: 10.1038/nbt1068.

[9] D. Frentz, C. A. B. Boucher, M. Assel, A. De Luca, M. Fabbiani, F. Incardona, P. Libin, N. Manca, V. Müller, B. O Nualláin, R. Paredes, M. Prosperi, E. Quiros-Roldan, L. Ruiz, P. M. A. Sloot, C. Torti, A. Vandamme, K. Van Laethem, M. Zazzi, and D. A. M. C. van de Vijver. Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time. *PLoS One*, 5(7):e11505, 2010. doi: 10.1371/journal.pone.0011505. URL `http://dx.doi.org/10.1371/journal.pone.0011505`.

[10] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02, 2004. Gaussian, Inc., Wallingford, CT, 2004.

[11] O. Gallego, L. Martin-Carbonero, J. Aguero, C. de Mendoza, A. Corral, and V. Soriano. Correlation between rules-based interpretation and virtual

phenotype interpretation of HIV-1 genotypes for predicting drug resistance in HIV-infected individuals. *J. Virol. Methods*, 121(1):115–118, Oct 2004. doi: 10.1016/j.jviromet.2004.06.003. URL `http://dx.doi.org/10.1016/j.jviromet.2004.06.003`.

[12] A.X. Garg, N.K.J. Adhikari, H. McDonald, M.P. Rosas-Arellano, P.J. Devereaux, J. Beyene, J. Sam, and R.B. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *J Am Med Ass*, 293:1223–1238, 2005.

[13] S. Genheden and U. Ryde. How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.*, 31(4):837–846, 2010. doi: 10.1002/jcc.21366.

[14] M. K. Gilson and H. Zhou. Calculation of protein-ligand binding affinities. *Ann. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007. doi: 10.1146/annurev.biophys.36.040306.132550.

[15] O. Guvench and A. D. MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Methods Mol. Biol.*, 443:63–88, 2008. doi: 10.1007/978-1-59745-177-2_4. URL `http://dx.doi.org/10.1007/978-1-59745-177-2\_4`.

[16] E. Jenwitheesuk and R. Samudrala. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct. Biol.*, 3:2, Apr 2003. doi: 10.1186/1472-6807-3-2.

[17] E. Jenwitheesuk and R. Samudrala. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir. Ther.*, 10(1):157–166, 2005.

[18] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983. doi: 10.1063/1.445869. URL `http://link.aip.org/link/?JCP/79/926/1`.

[19] R. K. Kancha, N. von Bubnoff, C. Peschel, and J. Duyster. Functional Analysis of Epidermal Growth Factor Receptor (EGFR) Mutations and Potential Implications for EGFR Targeted Therapy. *Clin. Cancer Res.*, 15:460–467, 2009. doi: 10.1158/1078-0432.CCR-08-1757.

[20] R. Kim and J. Skolnick. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.*, 29(8):1316–1331, Jun 2008. doi: 10.1002/jcc.20893. URL `http://dx.doi.org/10.1002/jcc.20893`.

[21] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, 2000. doi: 10.1021/ar000033j.

[22] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.*, 4(10):790–797, Oct 2006. doi: 10.1038/nrmicro1477.

[23] A. Luckow, S. Jha, J. Kim, A. Merzky, and B. Schnor. Adaptive distributed replica-exchange simulations. *Phil. Trans. R. Soc. A*, 367(1897):2595–2606, Jun 2009. doi: 10.1098/rsta.2009.0051. URL `http://dx.doi.org/10.1098/rsta.2009.0051`.

[24] J. Martinez-Picado, T. Wrin, S. D. W. Frost, B. Clotet, L. Ruiz, A. J. Brown, C. J. Petropoulos, and N. T. Parkin. Phenotypic hypersusceptibility to multiple protease inhibitors and low replicative capacity in patients who are chronically infected with human immunodeficiency virus type 1. *J. Virol.*, 79(10):5907–5913, May 2005. doi: 10.1128/JVI.79.10.5907-5913.2005. URL `http://dx.doi.org/10.1128/JVI.79.10.5907-5913.2005`.

[25] I. Massova and P.A. Kollman. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, 121(36):8133–8143, 1999. doi: 10.1021/ja990935j.

[26] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure*, 17(4):489–498, Apr 2009. doi: 10.1016/j.str.2009.02.010.

[27] H. Ohtaka, A. Schön, and E. Freire. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry*, 42(46):13659–13666, 2003. doi: 10.1021/bi0350405.

[28] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani. Global cancer statistics, 2002. *CA Cancer J. Clin.*, 55:74–108, 2005.

[29] A. L. Perryman, J. Lin, and J. A. McCammon. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci.*, 13 (4):1108–1123, 2004. doi: 110/ps.03468904.

[30] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005. doi: 10.1002/jcc.20289. URL `http://dx.doi.org/10.1002/jcc.20289`.

[31] J. Ravela, B. J. Betts, F. Brun-Vézinet, A. Vandamme, D. Descamps, K. van Laethem, K. Smith, J. M. Schapiro, D. L. Winslow, C. Reid, and R. W. Shafer. HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. *J. Acquir. Immune. Defic. Syndr.*, 33(1):8–14, May 2003.

[32] R. Rodriguez, G. Chinea, N. Lopez, T. Pons, and G. Vriend. Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, 14(6):523–528, 1998. doi: 10.1093/bioinformatics/14.6.523.

[33] S. K. Sadiq, M. D. Mazzeo, S. J. Zasada, S. Manos, I. Stoica, C. V. Gale, S. J. Watson, P. Kellam, S. Brew, and P. V. Coveney. Patient-specific simulation as a basis for clinical decision-making. *Phil. Trans. R. Soc. A*, 366(1878):3199–3219, 2008. doi: 10.1098/rsta.2008.0100. URL `http://dx.doi.org/10.1098/rsta.2008.0100`.

[34] S. K. Sadiq, D. Wright, S. J. Watson, S. J. Zasada, I. Stoica, and P.V. Coveney. Automated Molecular Simulation Based Binding Affinity Calculator for Ligand-Bound HIV-1 Proteases. *J. Chem. Inf. Model.*, 48(9):1909–1919, 2008. doi: 10.1021/ci8000937.

[35] S. K. Sadiq, D. W. Wright, O. A. Kenway, and P. V. Coveney. Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.*, 50(5):890–905, May 2010. doi: 10.1021/ci100007w.

[36] R. S. Saksena, B. Boghosian, L. Fazendeiro, O. A. Kenway, S. Manos, M. D. Mazzeo, S. K. Sadiq, J. L. Suter, D. Wright, and P. V. Coveney. Real science at the petascale. *Phil. Trans. R. Soc. A*, 367:2557–2571, 2009. doi: 10.1098/rsta.2009.0049.

[37] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996. doi: 10.1002/(SICI)1097-0282(199603)38:3. URL `http://dx.doi.org/gt;3.0.CO;2-Y`.

[38] C. Sawyers. Targeted cancer therapy. *Nature*, 432:294–297, 2004. doi: 10.1038/nature03095.

[39] A. W. Schüttelkopf and D. M. F. van Aalten. *PRODRG*: a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 60(8):1355–1363, 2004. doi: 10.1107/S0907444904011679. URL `http://dx.doi.org/10.1107/S0907444904011679`.

[40] M. D. Shenderovich, R. M. Kagan, P. N. R. Heseltine, and K. Ramnarayan. Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci.*, 12(8):1706–1718, Aug 2003. doi: 10.1110/ps.0301103. URL `http://dx.doi.org/10.1110/ps.0301103`.

[41] P. M. A. Sloot, P. V. Coveney, G. Ertaylan, V. Müller, C. A. Boucher, and M. Bubak. HIV decision support: from molecule to man. *Phil. Trans. R. Soc. A*, 367:2691–2703, 2009. doi: 10.1098/rsta.2009.0043.

[42] T. Steinbrecher and A. Labahn. Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies. *Curr. Med. Chem.*, 17:767–85, Jan 2010.

[43] I. Stoica, S.K. Sadiq, and P.V. Coveney. Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases. *J. Am. Chem. Soc.*, 130(8):2639–2648, 2008. ISSN 0002-7863. doi: 10.1021/ja0779250.

[44] S. Wan and P. V. Coveney. Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J. R. Soc. Interface*, 2011.

[45] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general Amber force field. *J. Comput. Chem.*, 25 (9):1157–1174, 2004. doi: 10.1002/jcc.20035. URL `http://dx.doi.org/10.1002/jcc.20035`.

[46] K. Wittayanarakul, S. Hannongbua, and M. Feig. Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J. Comput. Chem.*, 29(5):673–685, 2008. doi: 10.1002/jcc.20821.

[47] J. Yanchunas, D. R. Langley, L. Tao, R. E. Rose, J. Friborg, R. J. Colonno, and M. L. Doyle. Molecular basis for increased susceptibility of isolates with atazanavir resistance-conferring substitution I50L to other protease inhibitors. *Antimicrob. Agents Chemother.*, 49(9):3825–3832, Sep 2005. doi: 10.1128/AAC.49.9.3825-3832.2005. URL `10.1128/AAC.49.9.3825-3832.2005`.

[48] C. Yun, T.J. Boggon, Y. Li, M.S. Woo, H. Greulich, M. Meyerson, and M.J. Eck. Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. *Cancer Cell*, 11:217–227, 2007. doi: doi:10.1016/j.ccr.2006.12.017.

[49] C. Yun, K.E. Mengwasser, A.V. Toms, M.S. Woo, H. Greulich, K. Wong, M. Meyerson, and M.J. Eck. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Nat. Acad. Sci. USA*, 105:2070–2075, 2008. doi: 10.1073/pnas.0709662105.

[50] S. J. Zasada and P. V. Coveney. Virtualizing access to scientific applications with the Application Hosting Environment. *Comput. Phys. Commun.*, 180:2513–2525, 2009. doi: 10.1016/j.cpc.2009.06.008.

[51] S. J. Zasada, T. Wang, A. Haidara, E. Liub, N. Graf, G. Clapworthy, S. Manos, and P. V. Coveney. An e-Infrastructure Environment for Patient Specific Multiscale Modelling and Treatment. *Preprint submitted for publication*, 2010.

[52] H. Zhang, A. Berezov, Q. Wang, G. Zhang, J. Drebin, R. Murali, and M.I. Greene. ErbB receptors: from oncogenes to targeted cancer therapies. *J. Clin. Invest.*, 117:2051–2058, 2007. doi: 10.1172/JCI32278.

Supplementary Information: Towards Personalised Drug Ranking in Clinical Decision Support

By David W. Wright, Shunzhou Wan†, S. Kashif Sadiq, Stefan J. Zasada and Peter V. Coveney‡

Centre for Computational Science, Department of Chemistry,University College London, London, WC1H 0AJ, U.K.

Keywords: Molecular Dynamics, Patient Specific Medicine, Clinical Decision Support, HIV-1, Protease, Lopinavir, Cancer, EGFR, Gefitinib

The main article describes the application of an automated simulation workflow orchestrated by our Binding Affinity Calculator (BAC) tool to calculate binding free energies, from molecular dynamics simulations, for inhibitory drugs used in two distinct systems; the HIV-1 protease and the epidermal growth factor receptor (EGFR). This supporting information is elucidates the various steps of the simulation workflow, in particular the protocol used to equilibrate the system prior to the production phase from which free energies are computed.

1. Binding Affinity Calculator Workflow

Figure 1 shows the overall workflow of the Binding Affinity Calculator (BAC) as described in detail in Sadiq et al. [2008]. The first step is to produce a simulation-ready structure. This is generated from an initial set of coordinates, selected from a library of PDB structures in the BAC, together with the generic topology and forcefield parameter information. Suitable protease and ligand coordinates are extracted, any mutations incorporated, then charge neutralizing ions and water necessary for the solvation of the target structure are added. System-specific topology and coordinate files then have to be generated which form the input for subsequent simulation. The next stage involves the array of sequential equilibration simulations that need to run before production simulations can commence. These include the stages of minimization, annealing the system, the gradual relaxing of constraints which vary based on the mutations that have been incorporated and, finally, unrestrained equilibration in a specified thermodynamic ensemble. Once the simulation is complete the generated coordinate trajectories are post processed to calculate the enthalpic and entropic components of the binding free energy.

The steps involved in the equilibration protocol are adapted to account for the number of mutations inserted into the structure *in silico* before simulation and are described in greater detail below.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

‡ Corresponding author: p.v.coveney@ucl.ac.uk

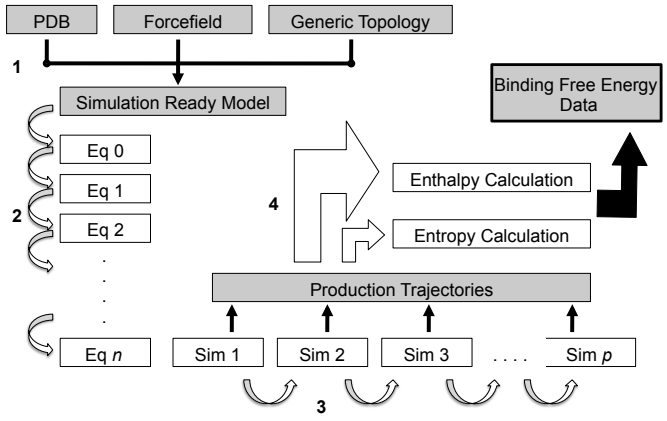


Figure 1: Workflow of an MMPBSA free energy calculation comprising four sequential stages. (1) Preparation of a simulation-ready model from the protein data bank crystal structure (PDB), forcefield parameters, and generic topology information. (2) Linear chain of equilibration simulations. (3) Linear chain of production simulations each generating trajectories for analysis. (4) Postproduction execution of the enthalpy and entropy calculations leading to a determination of the binding free energy. Data files are shown in gray boxes, processes, in white boxes. Adapted from Sadiq et al. [2008]

(*a*) *Mutation-Adaptive Equilibration Protocol*

A sequence of equilibration simulations are run prior to any BAC production simulation. Table 1 summarises each of these stages and their purposes. During the initial steps of equilibration the heavy atoms of the protein and ligand are constrained to their original positions. The mutation relaxation steps involve the removal of these constraints on all heavy atoms within 5Å of the mutated residue, the released residues being known as the 'M-region' during each step. Once the 50 ps relaxation step is complete constraints are reapplied to these residues (although they may be removed again if they fall within 5Å of a mutation being relaxed in the subsequent step). In dimeric systems (such as HIV-1 protease) where mutations at a given locus correspond to two positions in the three dimensional structure M-regions are constructed for both simultaneously. The mutation regions are selected in ascending numerical order of the mutated amino-acid residue number corresponding to the mutated locus. For example, if positions 48 and 90 are mutated, the first mutation region selected will contain any complete residues that are either partially or wholly within a 5 Å region around position 48 (potentially in both monomers of a dimeric system), while the second mutation region will be an identically defined region around positions 90 (again potentially in both monomers of a dimeric system). The length of the final step of the equilibration phase is adjusted to account for the number of relaxation steps so that the full equilibration phase lasts 2 ns in all cases.

			Force constraint (kcal/(mol Å^2))	
Stage	Process	Duration (ps)	Ligand	Protein
Eq 0	Minimisation	2000 steps	4	4
Eq 1	Annealing	50	4	4
Eq 2	NPT solvation	200	4	4
	Mutation Relaxation		M-region[a]	NM-region[b]
Eq (2 + 1)	M1-region relaxation	50	0	4
Eq (2 + 2)	M2-region relaxation	50	0	4
⋮	⋮	⋮	⋮	⋮
Eq (2 + n)	Mn-region relaxation	50	0	4
			Ligand	Protein
Eq (2 + n + 1)	Constraint removal (NPT)	50	3	4
Eq (2 + n + 2)		50	2	4
Eq (2 + n + 3)		50	1	4
Eq (2 + n + 4)		50	0	4
Eq (2 + n + 5)		50	0	3
Eq (2 + n + 6)		50	0	2
Eq (2 + n + 7)		50	0	1
Eq (2 + n + 8)	Unconstrained removal (NPT)	1400 - 50n	0	0

Table 1: The steps involved in the BAC equilibration protocol. [a]M-region consists of all heavy ligand or protein atoms within a 5 Å centred on each mutated residue (ligands are treated as a single residue). [b]NM-region consists of all heavy ligand or protein atoms outside the M-region.

References

S. K. Sadiq, D. Wright, S. J. Watson, S. J. Zasada, I. Stoica, and P.V. Coveney. Automated Molecular Simulation Based Binding Affinity Calculator for Ligand-Bound HIV-1 Proteases. *J. Chem. Inf. Model.*, 48(9):1909–1919, 2008. doi: 10.1021/ci8000937.