

## Supporting Information for

### Rapid and Reliable Binding Affinity Prediction of Bromodomain Inhibitors: a Computational Study

Shunzhou Wan,<sup>a</sup> Agastya P. Bhati,<sup>a</sup> Stefan J. Zasada,<sup>a</sup> Ian Wall,<sup>b</sup> Darren Green,<sup>b</sup> Paul Bamborough,<sup>b</sup> and Peter V. Coveney<sup>a\*</sup>

#### Affiliations:

<sup>a</sup>Centre for Computational Science, Department of Chemistry, University College London, London WC1H 0AJ, United Kingdom.

<sup>b</sup>GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, United Kingdom.

\*Correspondence to: [p.v.coveney@ucl.ac.uk](mailto:p.v.coveney@ucl.ac.uk)

The main article describes the prediction of binding free energies using the two approaches we have recently formulated, namely ESMACS (enhanced sampling of molecular dynamics with approximation of continuum solvent) and TIES (thermodynamic integration with enhanced sampling). The Supporting Information describes the approaches in more detail, together with, in order of appearance, decoupling of electrostatic and vdW interactions in the TIES study (Figure S1), additional analyses concerning the use of metadynamics to select the proper compound rotamers (Figure S2), and an ESMACS calculation including explicit water molecules (Figure S3).

## Materials and Methods

The detailed description of the ESMACS and TIES approaches has been provided in our recent manuscript. For the reader's convenience, we briefly describe these methods here.

**ESMACS protocol.** In ESMACS, the free energy is evaluated approximately based on the extended MMPBSA method, including configurational entropy, and the free energy of association. Our previous studies have shown that the sampled conformations are extensively enhanced in the ensemble simulations than these in single simulations.<sup>1-5</sup> According to the protocol we have recently developed, an ensemble consisting of 25 replica simulations is used for each complex. The replica simulations start from the same structures but use randomized initial velocities drawn from a Maxwell-Boltzmann distribution at 50 K. The systems are first heated in 1 K steps over a period of 60 ps to 300 K. This is followed by 2 ns equilibration followed by 4 ns production runs for each replica. All simulations are performed in an NPT ensemble using periodic boundary conditions. Trajectories are recorded every 10 ps in the production runs. The number of replicas and the simulation length of each replica are carefully determined<sup>4-5</sup> so that an optimal trade-off between precision and computational cost is achieved. This is followed by a free energy calculation at 50 uniformly chosen snapshots from each production run, leading to on the order of a thousand snapshots for each system. The NMODE calculations were performed for the entire complex/protein/ligand systems. Finally, data analysis is performed by a set of automated scripts

on a local workstation or remote machine yielding the distribution of free energies, that is the final reported  $\Delta G$ , with accompanying statistical error bars.

ESMACS can be performed using ensembles of 1-, 2- and/or 3-trajectory methods. In the 1-trajectory approach, the trajectories of the complexes, the receptors and the ligands are all extracted from the simulations of the complexes. In the 2-trajectory, the trajectories of the receptors or ligands are obtained from their own trajectories, while the trajectories of the complexes and the ligands/receptor (for which simulations are not performed) are extracted from the complex simulations. In the current study, the ranking of the ligand binding to a single protein is the main concern. An individual simulation for the protein provides a constant term in the binding free energy calculations (Equation 1 in the main text) for the 2- and 3-traj approaches. The average free energy of the protein from all of the complex simulations is therefore used when free energy of the apo protein is required. In the 3-trajectory approach, the trajectories of complexes, receptors and ligands are all obtained from their own separate simulations. It can prove highly informative to compare findings based on the 1-, 2- and 3-trajectory versions of ESMACS. For ligand-protein binding that approximates well the original lock-and-key hypothesis, 1-trajectory ESMACS will work well; for situations involving “induced fits” and other steric considerations, 2- and/or 3-trajectory ESMACS is required and is capable of providing many detailed insights into the binding energetics.

**TIES protocol.** The binding free energy difference,  $\Delta\Delta G_{\text{binding}}$ , of two ligands to the same protein, or a ligand to wild-type and mutant proteins can be computed using an alchemical transformation for the mutated entity in aqueous solution and within the ligand-protein complex. TIES by construction compares one congeneric ligand to another, so it is not easy to make assessments of the effect that an individual ligand has on the protein.

All TIES calculations are done using a dual topology for the system of interest. The dual topology contains hybrid groups of atoms which consist of atoms that will gradually vanish from the initial state, and that will gradually appear towards the final state. In the current study, hybrid molecules are constructed in which two compounds are present. A coupling parameter  $\lambda$  ( $0 \leq \lambda \leq 1$ ) is introduced into the potential function

$$V(\lambda) = (1 - \lambda)V_0(1 - \lambda) + \lambda V_1(\lambda) \quad (\text{S1})$$

to describe an intermediate state  $\lambda$ .  $\lambda=0$  and  $\lambda=1$  correspond to the initial and final thermodynamic states.

In the current TIES study, the alchemical coupling parameter (Equation S1) assumes the values  $\lambda=0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0$ . Additional potential energy values could be collected at each  $\lambda$  window for other post-processing analyses such as multistate Bennett acceptance ratio (MBAR)<sup>6</sup> so that the predictions from different approaches (TI or MBAR) can be compared. This is a feature is currently lacking in NAMD. The MBAR approach is often preferred over TI as it is said to exhibit less bias.<sup>7</sup> In a recent study,<sup>8</sup> however, it was found that the differences between results from two independent runs were much larger than those between results from the same run but using these two different approaches. This further emphasises the necessity for ensemble simulations to obtain reproducible predictions and clouds earlier claims of superiority.

As in ESMACS, an ensemble must be generated at each of the different intermediate values of  $\lambda$  in TIES to evaluate the energy derivatives numerically. The convergence analyses of our previous

studies show that 5 replicas and 4 ns simulation per replica at each  $\lambda$  window are sufficient to generate precise results, and are hence recommended for a TIES study. 2 ns equilibration and 4 ns production runs are therefore performed, as in the ESMACS study. As above, there is a trade-off to be struck between computational cost, accuracy and precision. Higher accuracy and precision can be obtained by increasing the number of replicas and/or the simulation length, the improvement scaling with the computational cost.

The difference between our approaches and others in the literature resides in the use of ensemble averaging and the recognition of the properties of Gaussian random processes computed from MD trajectories.<sup>1</sup> The criterion for ensuring convergence of the ensemble average is to establish the number  $N$  of replicas required such that using more replicas makes no significant difference to the values calculated. Although there are publications on free energy calculations using a few repeats of single molecular dynamics simulations,<sup>9-10</sup> only a very few systematic assessments have been made of the number of replicas required to produce converged results.<sup>5, 11-14</sup> Our primary goal in establishing the protocols has been to find a general computational way to rank the binding affinities without any *ad hoc* restrictions. In our previous publications,<sup>5, 14</sup> we provide very clear and specific answers to the size of ensembles required in a study. The protocols we provide in turn allow the calculations to be performed in a tractable way, and, moreover, on a time scale that is actionable. Our work demonstrates that time averages computed by “long” simulations (e.g. comprising more than at least five times the aggregate duration of all ensemble runs performed) typically produce behaviour no better than that of a single short time duration replica drawn from an ensemble.<sup>1, 4</sup>

**Standard error bars and convergence.** The accuracy, precision and reproducibility of the ESMACS and TIES protocols are achieved via ensemble averaging. All the components of the free energy, being derived from individual molecular dynamics trajectories, exhibit an intrinsic dynamical instability;<sup>1</sup> they each produce properties which are distributed as Gaussian random processes. Ensemble averaging ensures that we compute the macroscopic free energy with well defined mean and bounded standard error, and in the manner stipulated by statistical mechanics.

Using the data from all twenty-five replicas in ESMACS, we can compute the mean value of  $\Delta G$  which is our estimate of the macroscopic binding affinity. To do this, we use bootstrap resampling<sup>5</sup>, where the data from our sample of 25 values is resampled (with replacement) to get a bootstrap resample and the corresponding mean. A very large number of bootstrap samples (usually 100,000 in our studies) is generated, drawing with replacement from the population of the original data set. “Resample with replacement” allows some data in the original set to appear more than once, while others do not appear at all. The standard deviation of the means provides an estimate of the standard error in the calculated binding affinity.

In the case of the 3-trajectory method, the above mentioned bootstrap resampling is done in a slightly different manner. Instead of collecting the means of bootstrap resamples of  $\Delta G$  values, differences in the means of bootstrap resamples for all three individual  $G$  values (one each for the protein-molecule complex, the free protein and the free molecule) are calculated to get the bootstrap distribution of means. This is necessary since, in the 3-trajectory case, the individual  $G$  values are extracted from independent simulations whose values are uncorrelated with respect to the replica index (1 to 25).

The error analyses for TIES are done in a similar way to that in ESMACS. The standard errors are evaluated for each  $\lambda$  window using the bootstrap resampling as mentioned above. With five

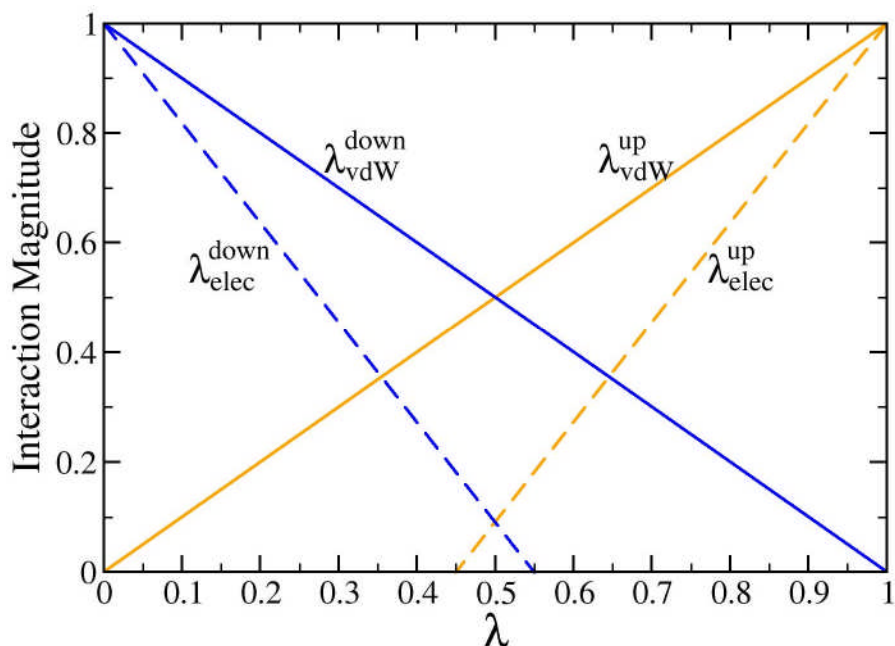
replicas employed in the current study, the standard errors in all of the TIES calculations are no greater than 0.2 kcal/mol.

The Spearman correlation coefficient is also calculated using bootstrap resampling<sup>11</sup>. All of reported binding free energies naturally have statistical errors. The correlations calculated from them are inevitably associated with uncertainties. For each experimental or calculated binding free energy, one new data point is generated by random sampling from its normal distribution with the reported mean and standard deviation. A correlation coefficient is then calculated for each of the newly generated dataset. A very large number of such bootstrap datasets (10,000,000 in this case in order to produce converged resample means) is constructed, from which the means and uncertainties are calculated.

**Implementation of the binding affinity calculator.** The binding affinity calculator (BAC) has already been used over several years to perform calculations on a range of high performance computers, including ones in the UK, the EU and the USA<sup>2-4;7;8</sup>. At the present time, the molecular simulation package NAMD<sup>9</sup> is used as the simulation engine, together with the Amber and AmberTools<sup>15</sup> for set up and analyses of the simulations. The simulation execution components of the BAC workflow are based on FabSim<sup>14</sup>, a Python-based automation toolkit for building scientific simulation and data processing workflows. This enables users to perform remote tasks from a local command-line, and to run applications while curating the environment variables along with the input and output data in a systematic manner.

We have recently developed a user friendly binding affinity calculator (*uf*-BAC) which enables BAC to be run via a Software as a Service model, hiding from the user the complexities of the command line tools used to build models, executing calculations on selected HPC resources, and analysing the results. *uf*-BAC plugs in to a range of computational back ends, from high performance computers provided by local, regional, national and international research facilities, to commercial cloud platforms. Users may create their own list of computing resources in a configuration file specifying their access privileges.

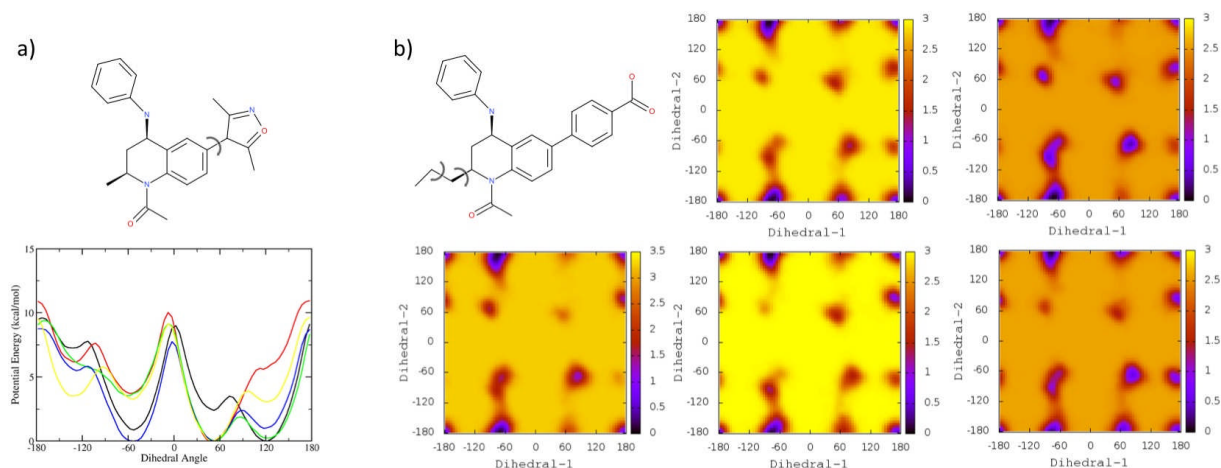
**Decoupling of electrostatic and van der Waals interactions.** In order to avoid “end-point catastrophes”,<sup>16</sup> a soft-core potential is used for pairwise vdW interactions involving the perturbed atoms. This is the reason why  $V_0$  and  $V_1$  in Equation S1 are also  $\lambda$ -dependent at the intermediate states. No soft-core potential is applied to the electrostatic interactions which are linearly scaled down for annihilating atoms and scaled up for growing atoms at a faster pace than the vdW interactions. The approach of decoupling at different paces removes the partial charges on perturbed atoms before they are fully annihilated, and introduces the charges on the growing atoms after they appear. In the current study, we set a user-defined value to be 0.55, such that a rate of  $\lambda_{elec}^{down} = 1/0.55 \times (0.55 - \lambda)$  can be used to scale down the electrostatic interactions of the annihilating atoms in the range of  $0 \leq \lambda \leq 0.55$ , and a rate of  $\lambda_{elec}^{up} = 1/0.55 \times (\lambda - 0.45)$  to scale up for the growing atoms in the range of  $0.45 \leq \lambda \leq 1$  (Figure S1). The  $\lambda_{elec}^{down}$  and  $\lambda_{elec}^{up}$  scale the electrostatic interactions down/up at faster paces than the corresponding  $\lambda_{vdW}^{down}$  and  $\lambda_{vdW}^{up}$  for the vdW interactions.



**Figure S1.** Decoupling van der Waals and electrostatic interactions at different paces. The electrostatic interactions are already scaled down to 0 at  $\lambda=0.55$  for the annihilating atoms which are completely annihilated at  $\lambda=1$ , and do not appear until  $\lambda>0.45$  for the growing atoms which are already presented extensively at  $\lambda=0.45$ .

**Metadynamics.** Metadynamics is a method that uses a history-dependent energy function to construct the potential energy landscape. The energy function, which changes and adapts over time, is then used to reconstruct a potential of mean force from a conventional molecular dynamics simulation.<sup>17</sup> Reaction coordinates, called collective variables, are used to represent progress along a reaction pathway. Here we used dihedral angles, the rotatable degrees of freedom, of the compounds as the collective variables since they determine the different binding poses of the compounds in the compound-protein complexes. 72 bins were used for each dihedral angle. When there were two dihedral angles involved, a uniform grid of  $72 \times 72$  was used. Five independent metadynamics simulations were run for those compounds which were expected to have more than one rotamer in their binding motif.

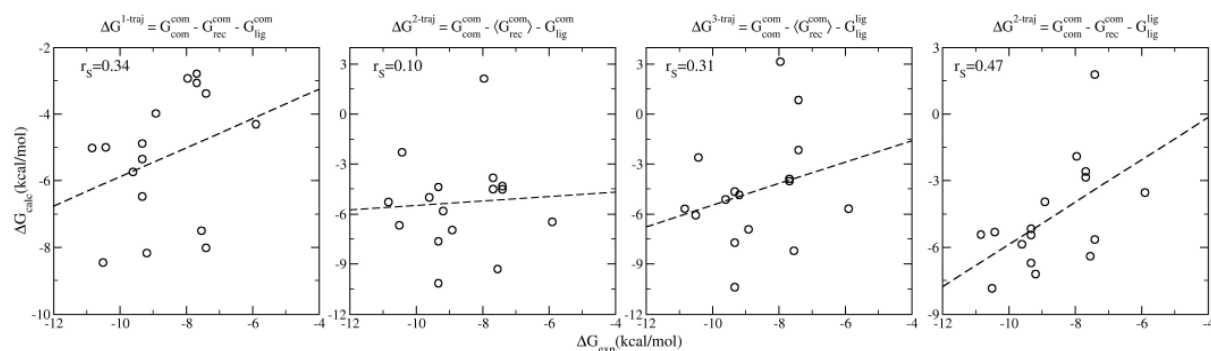
The reconstructed potential of mean force (see Figure S2) was then used to determine the binding pose(s) of a compound. As the metadynamics is used here to select the possible rotamers to set up the initial conformation for the ESMACS study, no attempt was made to fine tune the calculated potentials of mean force. The rotamer with the minimal potential of mean force was chosen as the initial conformation of the compounds in their complexes. For those with comparable minima, all possible rotamers were retained.



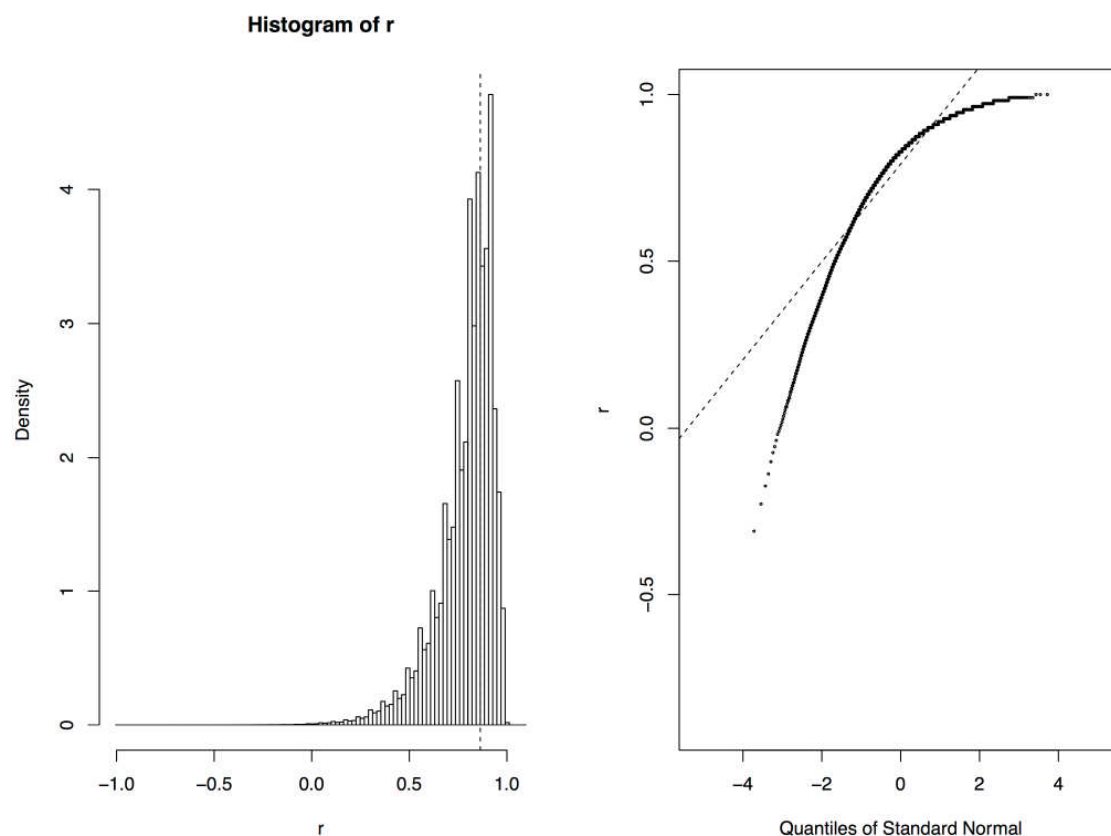
**Figure S2.** Potential of mean force reconstructed from metadynamics studies for compounds a) 7 and b) 13. Five independent simulations were performed. One dihedral angle was chosen as the collective variable for compound 7, and two dihedrals for compound 13. Three rotamers, with dihedral angle at  $-60^\circ$ ,  $60^\circ$  and  $120^\circ$ , were used for the ESMACS study of compound 7. One rotamer, with dihedral angles at  $\{-60^\circ, 180^\circ\}$  was used for compound 13.

## Additional Analyses

The four water molecules shown in Figure 1 of the main text may play an important role in ligand binding. Here we include them in the binding free energy calculation using the ESMACS approach. The water molecules may exchange with those from the surrounding solvent during the simulations. Four water molecules are extracted from each frame of the MD trajectories, which are located in the closest place to those in the X-ray structures. The ESMACS calculation is performed with the water molecules treated as part of the receptor. 1-traj ESMACS yields a weak correlation between the predictions and the experimental data, similar to that reported in the main text (Figure S3 and Figure 4). The 2-traj study, which takes into account the flexibility of the ligands, improves the predictions. Unlike the calculations without water molecules (Figure 4 in the main text), including the 4 water molecules as part of the receptor and treating its free energy as a constant significantly worsens the predictions. When ligands are absent, these water molecules frequently exchange with others from the bulk solvent, making it problematic to treat them as part of the receptor in the 2- and 3-traj approaches.



**Figure S3.** Spearman ranking correlations of the calculated binding free energies and the experimental data from 1-traj (left panel), 2-traj with constant receptor energy (centre left), 3-traj (centre right) and 2-traj with flexible ligands (right panel) ESMACS approaches. The equations on the top of each panel indicate the calculations used in each case. The subscripts (com/rec/lig) and the superscripts (com/lig) in the equations indicate the components (complexes, receptor and ligands) and the simulations (complexes and free ligands), respectively.



**Figure S4.** Non-normal distribution of Spearman ranking coefficients from bootstrap resampling. The tail towards the negative side makes the resampling mean of the coefficients smaller than that of the means of the binding free energy differences.

## References:

1. Coveney, P. V.; Wan, S., On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30236-30240.
2. Wan, S.; Coveney, P. V., Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J. R. Soc. Interface* **2011**, *8*, 1114-1127.
3. Wan, S.; Coveney, P. V., Rapid, accurate, and reproducible drug-protein binding affinity calculation. In *250th ACS National Meeting*, Boston, USA, 2015.
4. 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.* **2015**, *11*, 3346-3356.
5. Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V., Computing clinically relevant binding free energies of HIV-1 protease inhibitors. *J. Chem. Theory Comput.* **2014**, *10*, 1228-1241.
6. Shirts, M. R.; Chodera, J. D., Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.
7. Shirts, M. R.; Pande, V. S., Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **2005**, *122*, 144107.
8. Christ, C. D.; Fox, T., Accuracy assessment and automation of free energy calculations for drug design. *J. Chem. Inf. Model.* **2014**, *54*, 108-120.
9. Adler, M.; Beroza, P., Improved ligand binding energies derived from molecular dynamics: Replicate sampling enhances the search of conformational space. *J. Chem. Inf. Model.* **2013**, *53*, 2065-2072.
10. Lawrenz, M.; Baron, R.; McCammon, J. A., Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: Determinants of H5N1 avian influenza virus neuraminidase inhibition by peramivir. *J. Chem. Theory Comput.* **2009**, *5*, 1106-1116.
11. Genheden, S.; Ryde, U., How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **2010**, *31*, 837-846.
12. Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449-461.
13. Sadiq, S. K.; Wright, D. W.; Kenway, O. A.; Coveney, P. V., Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.* **2010**, *50*, 890-905.
14. Wan, S.; Coveney, P. V.; Flower, D. R., Peptide recognition by the T cell receptor: Comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations. *Phil. Trans A Math. Phys. Eng. Sci.* **2005**, *363*, 2037-2053.
15. Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.
16. Beveridge, D. L.; Dicapua, F. M., Free-energy via molecular simulation - applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431-492.
17. Cuendet, M. A.; Tuckerman, M. E., Free energy reconstruction from metadynamics or adiabatic free energy dynamics simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2975-2986.