PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2015, **17**, 8416

Evaluation of water displacement energetics in protein binding sites with grid cell theory†

G. Gerogiokas,[a] M. W. Y. Southey,[b] M. P. Mazanetz,[b] A. Hefeitz,[b] M. Bodkin,[b] R. J. Law[b] and J. Michel*[a]

Received 1st December 2014,
Accepted 8th January 2015

DOI: 10.1039/c4cp05572a

www.rsc.org/pccp

Excess free energies, enthalpies and entropies of water in protein binding sites were computed *via* classical simulations and Grid Cell Theory (GCT) analyses for three pairs of congeneric ligands in complex with the proteins scytalone dehydratase, p38α MAP kinase and EGFR kinase respectively. Comparative analysis is of interest since the binding modes for each ligand pair differ in the displacement of one binding site water molecule, but significant variations in relative binding affinities are observed. Protocols that vary in their use of restraints on protein and ligand atoms were compared to determine the influence of protein–ligand flexibility on computed water structure and energetics, and to assess protocols for routine analyses of protein–ligand complexes. The GCT-derived binding affinities correctly reproduce experimental trends, but the magnitude of the predicted changes in binding affinities is exaggerated with respect to results from a previous Monte Carlo Free Energy Perturbation study. Breakdown of the GCT water free energies into enthalpic and entropic components indicates that enthalpy changes dominate the observed variations in energetics. In EGFR kinase GCT analyses revealed that replacement of a pyrimidine by a cyanopyridine perturbs water energetics up three hydration shells away from the ligand.

Introduction

A long standing goal in computational chemistry is the routine accurate prediction of free energies of binding of drug-like small molecule ligands to proteins.[1] A strategic driver for this objective is its potential for significantly decreasing the resources and time commitments currently necessary for preclinical drug discovery activities.[2,3] A full description of protein–ligand interactions in aqueous environments requires a thorough analysis of the contributions of protein, ligand and solvent particles to the binding energetics. The role played by water in particular is the subject of intense research owing to its large influence on the binding process. In particular, there is extensive evidence that binding site water molecules are key players in this process.[4–7] This report focuses on the perturbations in binding site water network structure and associated energetics that occur upon small chemical modifications of small molecule ligands. This task is commonly attempted during hit-to-lead and lead optimisation phases of a structure-based drug discovery campaign.

Numerous computational methods have been developed to determine water location and energetics in binding sites owing to the difficulty of measuring these observables and quantities with experiments. A non-exhaustive list includes: the rolling probe-based Grid software; molecular dynamics probes based methods such as MDMix,[8] MixMD,[9] SILCS;[10] the Monte-Carlo λ-dynamics based algorithm JAWS,[11,12] inhomogeneous fluid solvation theory (IFST) based techniques,[13–16] including the popular method Watermap;[17] implicit and semi-explicit solvent methods such as SZMAP,[18] three dimensional reference interaction site model (3D-RISM),[19] and variational implicit solvent model (VISM).[20]

There is growing evidence that judicious use of the above methods is not only useful to further understanding of protein–ligand interactions with retrospective studies, but also to assist structure-based medicinal chemistry efforts. For instance the Watermap program was used at Pfizer to rationalise SAR and guide development of improved BACE-1 inhibitors.[21] In pursuit of a robust general methodology to this end our groups have recently proposed the Grid Cell Theory (GCT) methodology. The approach relies on a discretisation of the cell theory method developed by Henchman and co-workers,[22–25] to spatially resolve the free energy, enthalpy and entropy of water molecules at a protein interface. The methodology has been validated by prediction of the hydration thermodynamics of small molecules,[26] and has been applied to elucidate the binding thermodynamics of idealised host–guest systems.[27]

[a] *EaStCHEM School of Chemistry, Joseph Black Building, The King's Buildings, Edinburgh, EH9 3JJ, UK. E-mail: mail@julienmichel.net*
[b] *Evotec (UK) Limited, 114 Innovation Drive, Milton Park, Abingdon, Oxfordshire, OX14 4SA, UK*

† Electronic supplementary information (ESI) available: Additional figures on convergence of computed GCT energetics. See DOI: 10.1039/c4cp05572a

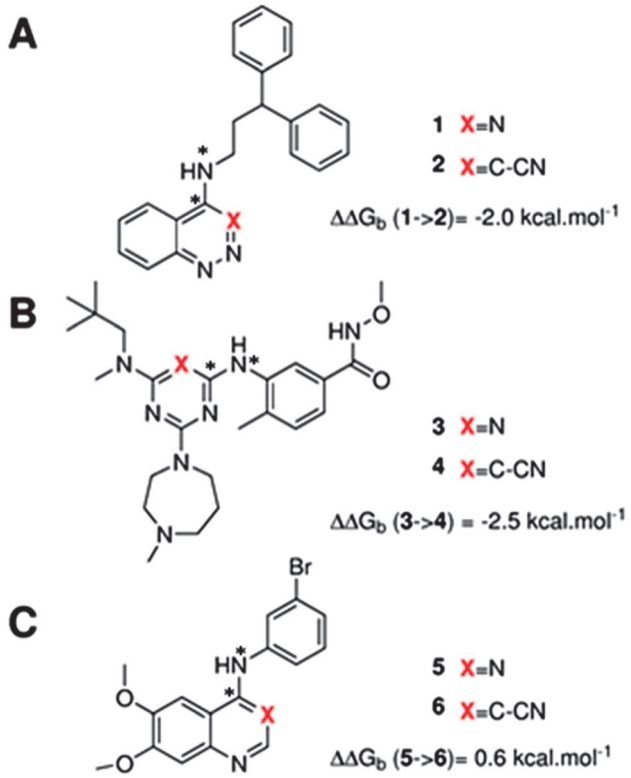

Fig. 1 Structures of the three pairs of ligands considered in this study. (A) Scytalone dehydratase, (B) p38a MAP kinase (C) EGFR kinase. Estimates of the experimental relative binding affinities are also shown. The star symbol denotes atoms used to define positional restraints (see Methods).

The GCT method is applied in the present report for the first time to protein–ligand complexes in order to elucidate the impact of ligand modifications on the thermodynamic properties of binding site water molecules. Pairs of congeneric ligands of three different proteins have been chosen for the present study, scytalone dehydratase,[28] p38α MAP kinase,[29] and EGFR kinase.[30] In each case, a similar strategy was used to displace a single binding site water molecule by introduction of a cyano group, yet significant differences were observed in the changes in binding affinity (Fig. 1). Previous computational work has reproduced the observed trends in relative binding affinities with the aid of Monte Carlo free energy perturbation methodologies (MC/FEP), but did not provide details of the enthalpic and entropic components of the binding affinities or the details of the water network perturbations.[31] The goals of the present study were, firstly to assess whether GCT is a competitive alternative to MC/FEP, secondly to determine solvent enthalpic and entropic contributions to binding affinities, and thirdly to determine the extent of binding site water perturbations upon a local ligand modification.

Theory and methods

Thermodynamic cycles

GCT analyses were performed on the basis of the thermodynamic cycle depicted in Fig. 2. The free energy change for water

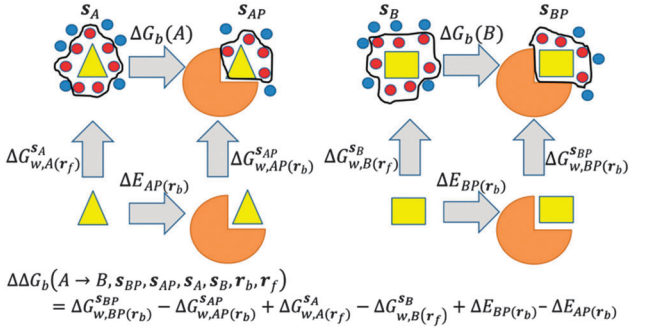

Fig. 2 Thermodynamic cycles for evaluation of water displacement free energies and relative free energies of binding. Ligands are depicted by yellow shapes. Proteins are depicted by orange shapes. In all GCT analyses, water molecules (red circles) inside the monitored regions $\boldsymbol{s}_A$, $\boldsymbol{s}_B$, $\boldsymbol{s}_{AP}$, $\boldsymbol{s}_{BP}$, contribute to the computed hydration free energies, whereas those that are out of the monitored regions are ignored (blue circles). Different restraint protocols $\boldsymbol{r}_c$ and $\boldsymbol{r}_l$ may be used to control allowed protein and ligand motions.

displacement is given by eqn (1):

$$\Delta\Delta G_{\mathrm{hyd}}(\mathrm{AP} \to \mathrm{BP}, \boldsymbol{s}_{\mathrm{BP}}, \boldsymbol{s}_{\mathrm{AP}}, \boldsymbol{r}_{\mathrm{c}}) = \Delta G^{\boldsymbol{s}_{\mathrm{BP}}}_{\mathrm{w,BP}(\boldsymbol{r}_{\mathrm{c}})} - \Delta G^{\boldsymbol{s}_{\mathrm{AP}}}_{\mathrm{w,AP}(\boldsymbol{r}_{\mathrm{c}})} \quad (1)$$

where $\Delta G^{\boldsymbol{s}_{\mathrm{AP}}}_{\mathrm{w,AP}(\boldsymbol{r}_{\mathrm{c}})}$ is the free energy of hydration of the monitored region $\boldsymbol{s}_{\mathrm{AP}}$ in the presence of ligand A and protein P with restraint protocol $\boldsymbol{r}_{\mathrm{c}}$, whereas expressions for ligand B are for the water displacing analogue. The water reorganisation free energy is given by eqn (2):

$$\begin{aligned}\Delta\Delta G_{\mathrm{water}}(\mathrm{A} \to \mathrm{B}, \boldsymbol{s}_{\mathrm{BP}}, \boldsymbol{s}_{\mathrm{AP}}, \boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}}, \boldsymbol{r}_{\mathrm{c}}, \boldsymbol{r}_{\mathrm{l}}) \\ = \Delta\Delta G_{\mathrm{hyd}}(\mathrm{AP} \to \mathrm{BP}, \boldsymbol{s}_{\mathrm{BP}}, \boldsymbol{s}_{\mathrm{AP}}, \boldsymbol{r}_{\mathrm{c}}) \\ - \Delta\Delta G_{\mathrm{hyd}}(\mathrm{A} \to \mathrm{B}, \boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}}, \boldsymbol{r}_{\mathrm{l}})\end{aligned} \quad (2)$$

where $\Delta\Delta G_{\mathrm{hyd}}(\mathrm{A} \to \mathrm{B}, \boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}}, \boldsymbol{r}_{\mathrm{l}})$ is the difference between the hydration free energy of ligands A $\left(\Delta G^{\boldsymbol{s}_{\mathrm{A}}}_{\mathrm{w,A}(\boldsymbol{r}_{\mathrm{l}})}\right)$ and B $\left(\Delta G^{\boldsymbol{s}_{\mathrm{B}}}_{\mathrm{w,B}(\boldsymbol{r}_{\mathrm{l}})}\right)$ computed using regions $\boldsymbol{s}_{\mathrm{A}}$ and $\boldsymbol{s}_{\mathrm{B}}$ and restraint protocol $\boldsymbol{r}_{\mathrm{l}}$. Relative free energies of binding are obtained with eqn (3):

$$\begin{aligned}\Delta\Delta G_{\mathrm{b}}(\mathrm{A} \to \mathrm{B}, \boldsymbol{s}_{\mathrm{BP}}, \boldsymbol{s}_{\mathrm{AP}}, \boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}}, \boldsymbol{r}_{\mathrm{c}}, \boldsymbol{r}_{\mathrm{l}}) \\ = \Delta\Delta G_{\mathrm{water}}(\mathrm{A} \to \mathrm{B}, \boldsymbol{s}_{\mathrm{BP}}, \boldsymbol{s}_{\mathrm{AP}}, \boldsymbol{s}_{\mathrm{A}}, \boldsymbol{s}_{\mathrm{B}}, \boldsymbol{r}_{\mathrm{c}}, \boldsymbol{r}_{\mathrm{l}}) \\ + \Delta\Delta E(\mathrm{AP} \to \mathrm{BP}, \boldsymbol{r}_{\mathrm{c}})\end{aligned} \quad (3)$$

where $\Delta\Delta E(\mathrm{AP} \to \mathrm{BP}, \boldsymbol{r}_{\mathrm{c}})$ is difference of the interaction energy of ligands A ($\Delta E_{\mathrm{AP}}(\boldsymbol{r}_{\mathrm{c}})$) and (B $\Delta E_{\mathrm{BP}}(\boldsymbol{r}_{\mathrm{c}})$) with the protein P. Contributions from relative changes in ligands internal energies, translational/rotational entropies, and ligand–protein conformational entropies are neglected in the present cycle. The approximation is only expected to be reasonable for comparisons of congeneric ligands that adopt the same binding mode.

Grid cell theory

Hydration free energies were computed using the grid cell theory method. In this approach the density, enthalpy, entropy and free energy of water are evaluated for an arbitrary region of

space $\boldsymbol{s}$ surrounding a solute X restrained by protocol $\boldsymbol{r}$. The approach involves three steps.

• Evaluation of cell parameters for water molecules within $\boldsymbol{s}$:

For each frame f to analyse, cell parameters for each of the N_f water molecules $i \in \boldsymbol{s}$ are determined. The cell parameters are: the solute–water interaction energy $\Delta H_i^{X(\boldsymbol{r})}$; the water–water interaction energy ΔH_i^{w}; the principal axes components of the force F_i^j ($j = x, y, z$) acting on the water molecule; the principal axes components of the torques τ_i^j ($j = x, y, z$) acting on the water molecule; the orientational number Ω_i^{ori} of the water molecule. Detailed expressions for these quantities are available in ref. 26.

• Evaluation of voxels parameters within $\boldsymbol{s}$:

The region $\boldsymbol{s}$ is decomposed into N_{s} cubic voxels of volume $V(k)$. Cell parameters are computed for each voxel k according to eqn (4):

$$A(k) = \frac{\sum_{f=1}^{M}\sum_{i=1}^{N_f} A_i I_k(i)}{\max\left\{1, \sum_{f=1}^{M}\sum_{i=1}^{N_f} I_k(i)\right\}} \quad (4)$$

where $A = \Delta H^{X(\boldsymbol{r})}$, ΔH^{w}, F^j, τ^j, Ω^{ori}. $I_k(i)$ is an indicator function that is equal to 1 if water molecule i is in voxel k, and 0 otherwise. M is the number of frames analysed. The average number of water molecules per voxel k is given by eqn (5).

$$N_{\mathrm{w}}(k) = \frac{1}{M}\sum_{f=1}^{M}\sum_{i=1}^{N_f} I_k(i) \quad (5)$$

• Evaluation of thermodynamic properties for $\boldsymbol{s}$:

Solute and solvent components of the enthalpy of hydration of region $\boldsymbol{s}$ are given by eqn (6) and (7):

$$\Delta H^{s}_{X(\boldsymbol{r})} = \sum_{k=1}^{N_{\mathrm{s}}} N_{\mathrm{w}}(k)\Delta H^{X(\boldsymbol{r})}(k) \quad (6)$$

$$\Delta H^{\boldsymbol{s}}_{\mathrm{w}} = \sum_{k=1}^{N_{\mathrm{s}}} N_{\mathrm{w}}(k)\Delta H^{\mathrm{w}}(k) \quad (7)$$

The excess enthalpy of water in region $\boldsymbol{s}$ is given by eqn (8):

$$\Delta H^{\boldsymbol{s}}_{\mathrm{w},X(\boldsymbol{r})} = \Delta H^{\boldsymbol{s}}_{X(\boldsymbol{r})} + \Delta H^{\boldsymbol{s}}_{\mathrm{w}} \quad (8)$$

Noting that the average number of water molecules within $\boldsymbol{s}$ is:

$$N_{\mathrm{w}}(\boldsymbol{s}) = \sum_{k=1}^{N_{\mathrm{s}}} N_{\mathrm{w}}(k) \quad (9)$$

Expressions for the average orientational number and forces/torques of region $\boldsymbol{s}$ are given by eqn (10):

$$A(\boldsymbol{s}) = \frac{1}{N_{\mathrm{w}}(\boldsymbol{s})}\sum_{k=1}^{N_{\mathrm{s}}} N_{\mathrm{w}}(k)A(k) \quad (10)$$

where $A = \Omega^{\mathrm{ori}}$, F^j, τ^j, noting that the minimum value for $\Omega^{\mathrm{ori}}(\boldsymbol{s})$ is always 1 in this study. Entropic components are given by eqn (11)–(13):

$$\Delta S^{\boldsymbol{s},\mathrm{ori}}_{\mathrm{w},X(\boldsymbol{r})} = N_{\mathrm{w}}(\boldsymbol{s})k_{\mathrm{B}}\ln\left\{\frac{\Omega^{\mathrm{ori}}(\boldsymbol{s})}{\Omega^{\mathrm{ori}}(\mathrm{bulk})}\right\} \quad (11)$$

$$\Delta S^{\boldsymbol{s},\mathrm{vib}}_{\mathrm{w},X(\boldsymbol{r})} = N_{\mathrm{w}}(\boldsymbol{s})k_{\mathrm{B}}\ln\left\{\prod_{j=1}^{3}\frac{F^j(\mathrm{bulk})}{F^j(\boldsymbol{s})}\right\} \quad (12)$$

$$\Delta S^{\boldsymbol{s},\mathrm{lib}}_{\mathrm{w},X(\boldsymbol{r})} = N_{\mathrm{w}}(\boldsymbol{s})k_{\mathrm{B}}\ln\left\{\prod_{j=1}^{3}\frac{\tau^j(\mathrm{bulk})}{\tau^j(\boldsymbol{s})}\right\} \quad (13)$$

where Ω^{ori}(bulk), F^j(bulk), τ^j(bulk) ($j = x, y, z$) are the cell parameters for the simulated water model in bulk conditions. The excess entropy of water within region $\boldsymbol{s}$ is given by eqn (14):

$$\Delta S^{\boldsymbol{s}}_{\mathrm{w},X(\boldsymbol{r})} = \Delta S^{\boldsymbol{s},\mathrm{ori}}_{\mathrm{w},X(\boldsymbol{r})} + \Delta S^{\boldsymbol{s},\mathrm{vib}}_{\mathrm{w},X(\boldsymbol{r})} + \Delta S^{\boldsymbol{s},\mathrm{lib}}_{\mathrm{w},X(\boldsymbol{r})} \quad (14)$$

The excess free energy of water is obtained from the enthalpic and entropic components:

$$\Delta G^{\boldsymbol{s}}_{\mathrm{w},X(\boldsymbol{r})} = \Delta H^{\boldsymbol{s}}_{\mathrm{w},X(\boldsymbol{r})} - T\Delta S^{\boldsymbol{s}}_{\mathrm{w},X(\boldsymbol{r})} \quad (15)$$

Eqn (4)–(15) were evaluated with the trajectory analysis software Nautilus. Details of the potential energy functions, regions and restraint protocols used in the present study are given below.

Restraints protocols

GCT calculations were performed with several different protocols that vary in their use of restraints to control the conformations sampled by the ligands or protein during the simulations. GCT calculations can in principle be performed without any restraints on solutes; however this has a number of disadvantages. Firstly, extensive conformational sampling is required to obtain converged water properties for flexible solutes. Secondly, graphical analyses of voxel properties are more complex. Thirdly, the thermodynamic cycle depicted in Fig. 2 doesn't consider contributions from changes in conformations or flexibility from the protein and ligands. On the other hand restraints are artificial and may negatively affect the predictions of free energies of binding. In the present work different restraining protocols $\boldsymbol{r}$ were compared in an effort to identify a practical protocol for routine calculations.

In the $\boldsymbol{r}$ = rot protocol positional restraints were applied on two atoms of a ligand. This was done to suppress rigid body motions of the ligand. For all ligands the restrained atom is denoted by a star in Fig. 1. In the $\boldsymbol{r}$ = bb protocol, positional restraints were applied to protein backbone heavy atoms only. Finally, in the $\boldsymbol{r}$ = full protocol, positional restraints were applied to all heavy atoms of both ligand and protein. When in solution with the full protocol, ligands were restrained in their binding site conformation. Restraints were implemented with a force constant of 10 kcal mol^{-1} Å^{-2} for the bb and full protocols, and with a force constant of 5 kcal mol^{-1} Å^{-2} for the rot protocol. All restraints were applied on absolute Cartesian coordinates.

Preparation of molecular models

Models of scytalone dehydratase in complex **1** and **2** were generated using the PDB structure 3STD which was in complex

with **2**. The crystal structure of EGFR kinase in complex with erlotinib (PDB 1M17) was used to define the binding mode of **5** and **6**. For p38 MAP kinase case the crystal structure of PDB 1DI9 which is in complex with a quinazoline inhibitor, was used to generate the protein model. AutoDock Vina[32] with the pymol plugin[33] was used to find suitable binding modes for **3** and **4** that matched structural data reported by Liu *et al.*[29] for **4**. The lowest energy pose produced by Vina for **4** was found to bind in a similar orientation.

The TIP4P-Ew water model was used throughout.[34] All the small molecules were parameterized using the GAFF force field[35] and AM1-BCC charges,[36] as implemented in the AMBER11 software suite.[37] For the protein, the ff12SB force field was used. Each protein complex and ligand was solvated with water extending 12 angstrom away from the edge of the solutes before performing energy minimisation. The preparation of molecular models was largely automated by the use of the software FESetup.[38]

Molecular dynamics simulations

Molecular simulations were produced using the software Sire/OpenMM wich in the present study results from linking of the general purpose molecular simulation package Sire revision 1786, with the GPU molecular dynamics library OpenMM revision 3537.[39] Simulations were run at 1 atm and 298 K using an atom-based generalized reaction field nonbonded cutoff of 10 Å for the electrostatic interactions,[40] and an atom-based non bonded cutoff of 10 Å for the Lennard-Jones interactions. A velocity-Verlet integrator with a time step of 2 fs was used. Temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1}.[41] Pressure control used isotropic box edge scaling Monte Carlo moves every 25 time steps. The OpenMM default error tolerance settings were used to constrain the intramolecular degrees of freedom of water molecules. For each system three simulations of 22 ns were run using the same starting conformation but a different random velocity assignment. Snapshots were stored every 1 ps and were written into a DCD format. The first 1 ns of each trajectory was discarded to enable equilibration.

Grid cell theory analyses

All GCT analyses were performed with the trajectory post-processing software Nautilus.[26] Bulk parameters for TIP4P-Ew were taken from a previous GCT study.[26] The following protocols were used to define the regions of space subjected to Nautilus analyses. For each simulation, a 3D grid of evenly spaced points was centered on the Cartesian coordinates of the centre of mass of the ligand (x_{com}, y_{com}, z_{com}). A rectangular region with minimum and maximum coordinates ($x_{com} \pm \Delta x$, $y_{com} \pm \Delta y$, $z_{com} \pm \Delta z$) was next defined and filled with grid points spaced every 0.5 Å along the x, y, and z components. The parameters Δx, Δy and Δz were chosen such that the grid would extend well beyond the ligand atoms or binding site region of interest (typical values are 11–14 Å). Cell parameters for every grid point within this rectangular region were then computed. For the simulations of the unbound ligands with the restraint protocol $\boldsymbol{r}$ = full, regions $\boldsymbol{s}_A/\boldsymbol{s}_B$ were defined as the union of the set of grid points that were within X_{vdw} Å of the van-der-Waals surface of the ligands A or B respectively. AMBER GAFF force-field radii were used to define the van-der-Waals surface from the input ligand coordinates and several values of X_{vdw} were tested. For the simulations of the unbound ligands with the protocol $\boldsymbol{r}$ = rot, regions $\boldsymbol{s}_A/\boldsymbol{s}_B$ were defined as the length X_{cubic} of the edge of a cube centred on (x_{com}, y_{com}, z_{com}).

For the simulations of the bound ligands with the restraint protocol $\boldsymbol{r}$ = full or $\boldsymbol{r}$ = bb, regions $\boldsymbol{s}_{AP}$, $\boldsymbol{s}_{BP}$ were defined by density-clustering of the trajectories of AP and BP. All grid points were first sorted by their local water density $\rho(k)$ in the simulation of AP. The medoid of a cluster was taken to be the grid point with highest density, and all grid points within 1.5 Å of this medoid were assigned to the cluster. All grid points belonging to the cluster were then removed from the grid and the process was iterated until no grid point with a density greater than 1.5 times bulk was found, yielding k_{AP} medoids. The process was repeated for the trajectory of BP, yielding k_{BP} medoids. Next, only medoids present in the binding site region of interest were retained, these were typically medoids present in binding site regions disconnected from bulk.

Next regions $\boldsymbol{s}_{AP}$ or $\boldsymbol{s}_{BP}$ were defined by selecting all grid points within X_{medoid} Å of each of the k_{AP} and k_{BP} medoids In some instances, the medoids from the AP or BP simulations had very similar coordinates and a single medoid was retained. The procedure yielded a monitoring region $\boldsymbol{s}_C$ that is the union of $\boldsymbol{s}_{AP}$ and $\boldsymbol{s}_{BP}$. In some instances, additional analyses were performed by breaking-down $\boldsymbol{s}_{AP}$ or $\boldsymbol{s}_{BP}$ into M sub-regions $\{\boldsymbol{s}_0, \ldots, \boldsymbol{s}_m\}$. This was done by defining a centre $\boldsymbol{r}_m$ = (x_m, y_m, z_m) for each of the M regions. The distance d_{im} of each grid point i in $\boldsymbol{s}_{AP}/\boldsymbol{s}_{BP}$ to each $\boldsymbol{r}_m$ was computed and the grid point i was assigned to the region M with the smallest value of d_{im}.

Results

Ligand hydration energetics

Fig. 3A shows that the computed free energy of hydration of a ligand in the GCT formalism depends on the size of the monitored region. With the $\boldsymbol{r}$ = full restraint protocol, hydration free energies have converged for regions that extend approximately 6 Å away from the van der Waals surface of the ligands. For regions of this size, uncertainties in the absolute hydration free energies are on the order of 1 kcal mol^{-1}. As discussed elsewhere, GCT hydration free energies are less precise than those computed by FEP or TI approaches because of the contribution of water–water interaction energies to the enthalpy of hydration (Fig. 3B). The entropies of hydration (Fig. 3C) are by contrast typically slightly better converged.[25] For the purpose of computing relative binding free energies between a pair of ligands, relative free energies of hydration are sufficient, and Fig. 3A shows that reasonable estimates of the difference can be estimated with smaller regions that extend about 4 Å away from the van der Waals surface of the ligands. The use of very large GCT regions is actually detrimental to accuracy since the

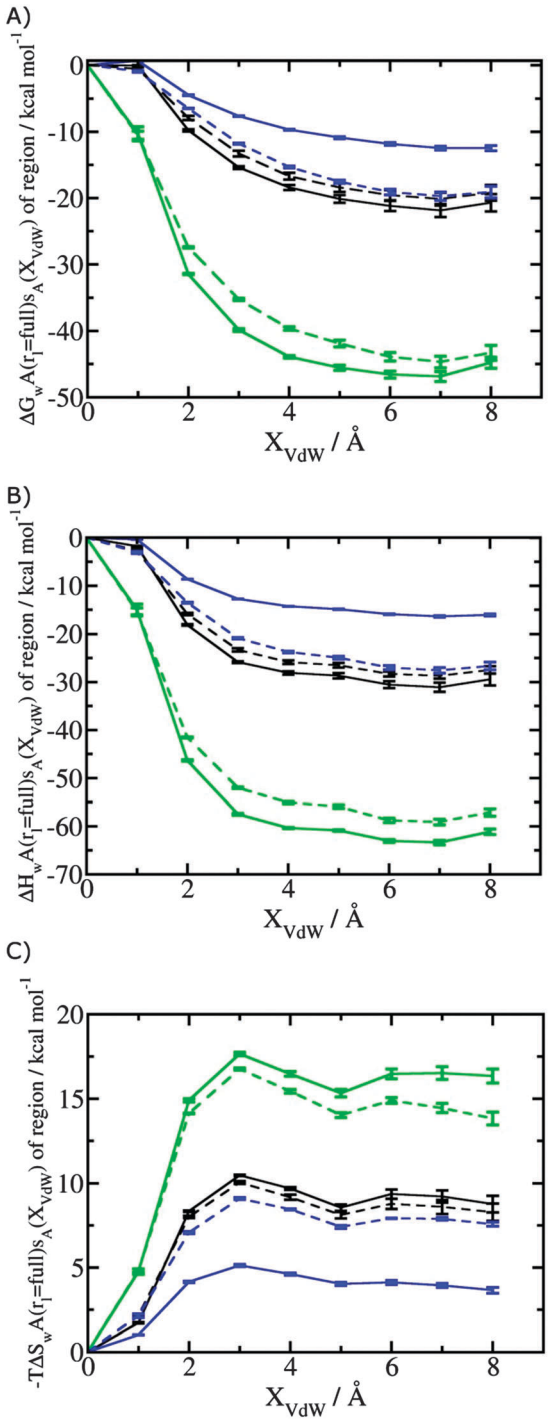

Fig. 3 Ligand hydration energetics with the full restraints protocol (A) hydration free energies $\Delta G^{s_A}_{w,A(r_l)}$, (B) hydration enthalpies $\Delta H^{s_A}_{w,A(r_l)}$, and (C) hydration entropies $-T\Delta S^{s_A}_{w,A(r_l)}$ black lines are for the scytalone dehydratase ligands **1** (solid) and **2** (dashed). Green lines are for the p38 MAP kinase ligands **3** (solid) and **4** (dashed). Blue lines are for the EGFR kinase ligands **5** (solid) and **6** (dashed). Error bars represent the standard error of the mean computed from triplicate independent simulations.

magnitude of uncertainties increases with the size of the region. Overall for these ligands, a good trade-off is to select a value of the parameter X_{VdW} between 4–6 Å.

Fig. 4 shows the computed hydration free energies (Fig. 4A), hydration enthalpies (Fig. 4B), and hydration entropies (Fig. 4C) with the r_l = rot protocol. Convergence of the computed hydration energetics is observed for cubes of edge length X_{cubic} *ca.* 10 Å. The uncertainties in the computed quantities are larger than with r_l = full protocol since the volume of the monitored region is actually larger. The relative hydration free energies are broadly comparable between the two restraint protocols for ligands **1**, **2** and for ligands **3**, **4**, but a noticeable discrepancy is apparent for ligands **5**, **6** $\Delta\Delta G_{hyd}$(**5** → **6**, s_6, s_5, r_l = full, X_{VdW} = 6 Å) = −7.3 ± 0.5 kcal mol^{-1}, *versus* $\Delta\Delta G_{hyd}$(**5** → **6**, s_6, s_5, r_l = rot, X_{cubic} = 9 Å) = −0.3 ± 1.5 kcal mol^{-1}. Visualization of the trajectories indicates that this likely occurred because the pyrimidine N1 nitrogen of **5** is poorly hydrated owing to the close proximity of the bromophenyl group in the r_l = full simulations. This occurred because the ligand was restrained to adopt the binding mode seen in the complex with EGFR kinase. Without such restraints in the r_l = rot simulations, **5** relaxed to a different conformation that increases hydration of the pyrimidine N1 nitrogen in **5**.

Protein–ligand complex hydration energetics

Fig. 5A shows the convergence of $\Delta\Delta G_{hyd}$(AP → BP, s_{BP}, s_{AP}, r_c) as a function of time, for three different monitored regions s_c defined by varying the parameter X_{medoid}, and for two different restraining protocols r_c. For low or intermediate values of X_{medoid}, similar results are obtained and trajectories of *ca.* 15 ns are needed to observe convergence. The same hydration free energy is obtained because the larger region defined with X_{medoid} = 4 Å still includes only one water molecule. However for X_{medoid} = 8 Å, the hydration free energies differ markedly because the monitored region s_c is now sufficiently large that it includes additional water molecules, some of them located out of the binding site of scytalone dehydratase. The hydration energetics are therefore different, and in the case of the r_c = bb protocol, no convergence is observed. The hydration energies between the r_c = bb and r_c = full protocols are not consistent because conformational changes in protein residues during the r_c = bb simulations affect the energetics of water molecules within the monitored region s_c. Similar variability is seen for p38 MAP kinase and EGFR kinase (Fig. S1 in ESI†).

Fig. 5B shows the computed hydration energetics for the three complexes using the full trajectories, but varying X_{medoid}. The plots show that the changes in hydration energetics between ligand pairs is relatively constant for small values of X_{medoid}, and the r_c = full protocol. Larger fluctuations are seen for EGFR kinase since the monitoring region s_c is larger and contains more water molecules. Larger values of X_{medoid}, or the additional protein flexibility in the r_c = bb protocol, causes increased statistical errors and fluctuations in the computed energetics. This indicates that much longer trajectories would be needed to obtain well reproducible changes in hydration energetics of the complexes. Consequently similar variability is seen in the evaluation of water reorganisation energies with eqn (3) (Fig. 5C). Overall, with trajectories of the order of *ca.* 10 ns, it

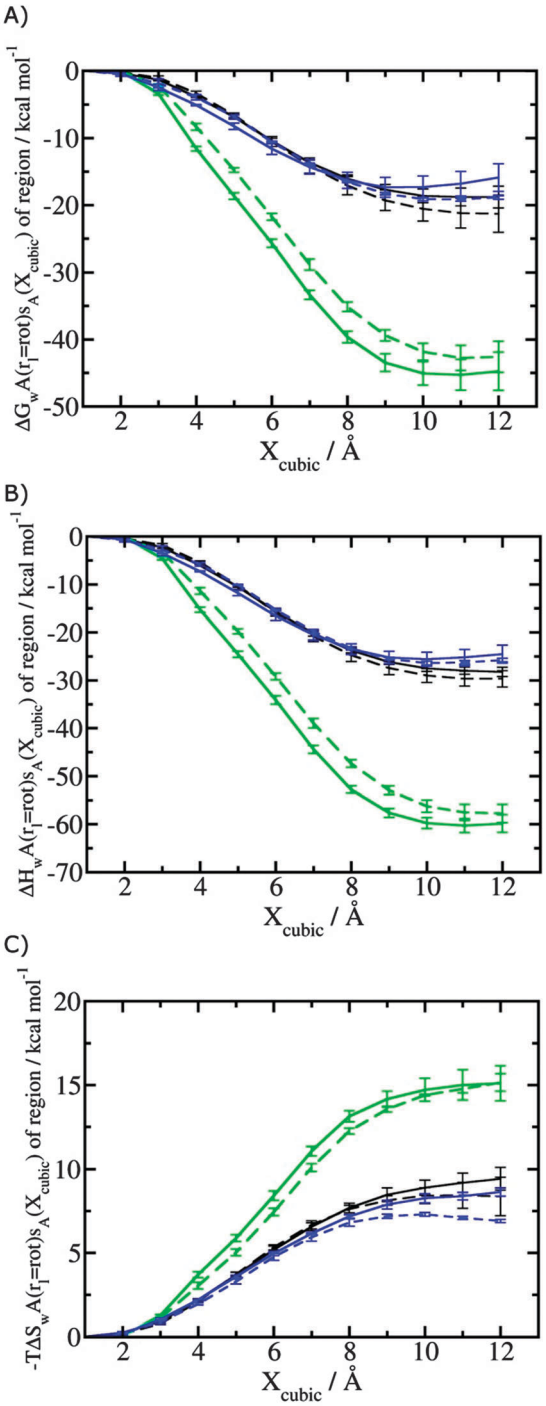

Fig. 4 Ligand hydration energetics with the rigid body rotation restraints protocol (A) hydration free energies $\Delta G^{s_A}_{w,A(r_l)}$, (B) hydration enthalpies $\Delta H^{s_A}_{w,A(r_l)}$, and (C) hydration entropies $-T\Delta S^{s_A}_{w,A(r_l)}$ black lines are for the scytalone dehydratase ligands **1** (solid) and **2** (dashed). Green lines are for the p38 MAP kinase ligands **3** (solid) and **4** (dashed). Blue lines are for the EGFR kinase ligands **5** (solid) and **6** (dashed). Error bars represent the standard error of the mean computed from triplicate independent simulations.

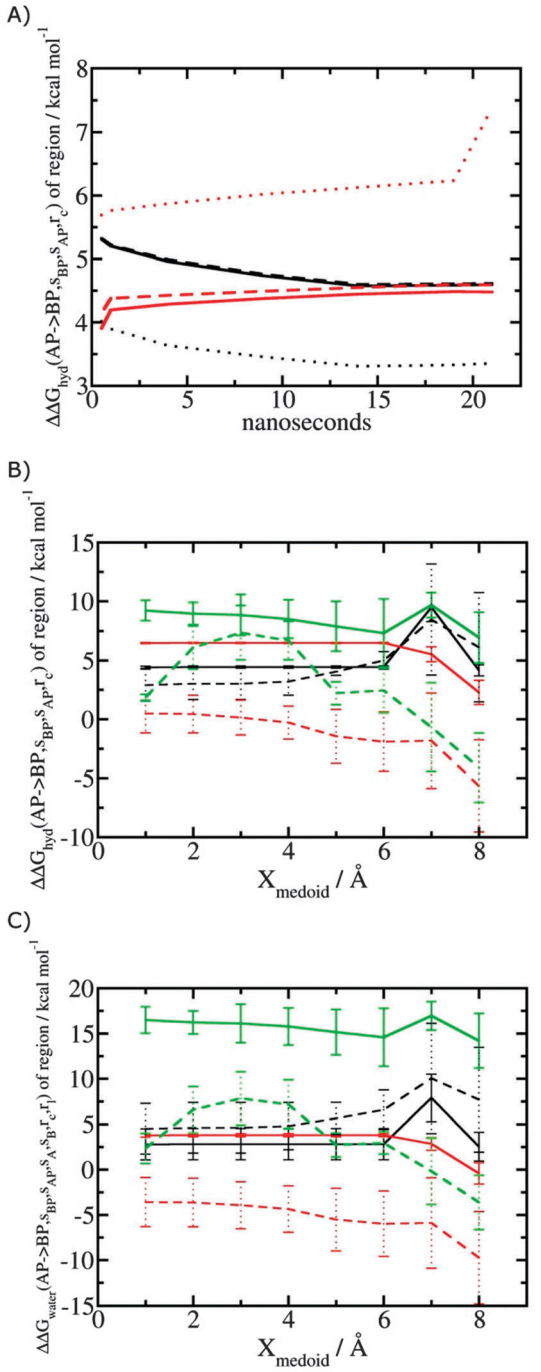

Fig. 5 Convergence of hydration energetics and water reorganisation energetics for protein–ligand complexes. (A) Convergence of hydration energetics (eqn (1)) with respect to trajectory duration for scytalone dehydratase. Results in black are for r_c = full, and in red for r_c = bb. The solid line is for X_{medoid} = 1 Å, the dashed line for X_{medoid} = 4 Å, and the dotted line for X_{medoid} = 8 Å. (B) Hydration energetics as a function of X_{medoid} using the full trajectories for scytalone dehydratase (black) p38 MAP kinase (red), and EGFR kinase (green). Solid lines are the results obtained with the r_c = full protocol and dotted lines are the results obtained with the r_c = bb protocol. (C) Same as (B) but for the water reorganisation energy (eqn (2)) using r_l = full, X_{VdW} = 6 Å or r_l = rot, X_{cubic} = 10 Å.

seems advisable to use the r_c = full protocol with X_{medoid} values between 4 to 6 Å if reproducible hydration energies are desired.

Fig. 6A shows the water content of the monitored region for the scytalone dehydratase/**1** complex. A single buried water molecule is present, hydrogen-bonded to two nearby tyrosine side-chains, and the nitrogen N1 of **1**. As expected, the water molecule is displaced in the scytalone dehydratase/**2** complex,

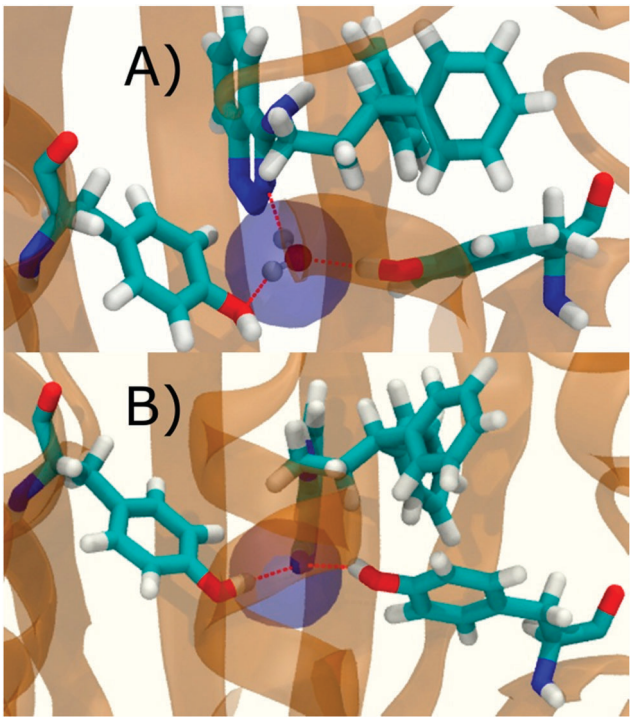

Fig. 6 Representation of GCT monitored regions in scytalone dehydratase. (A) In complex with **1** (B) in complex with **2**. Regions s_{AP}, and s_{BP} are depicted by the transparent blue spheres for r_c = full and X_{medoid} = 4 Å. The regions obtained with r_c = bb and X_{medoid} = 4 Å conditions are not shown because they are similar. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

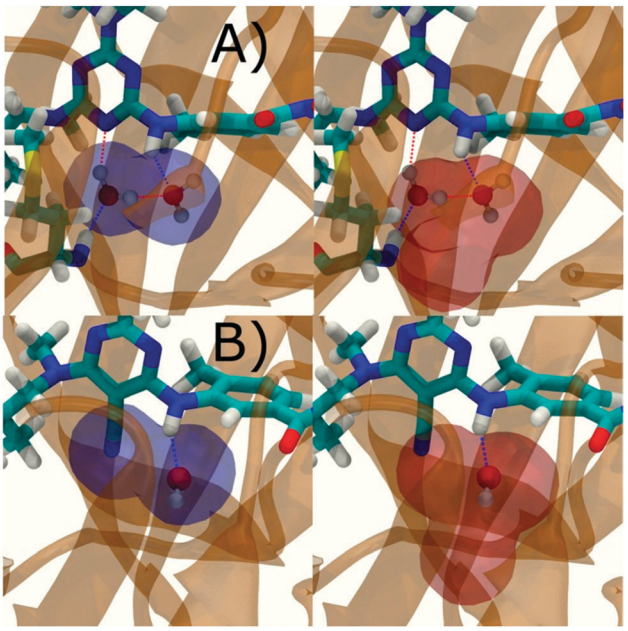

Fig. 7 Representation of GCT monitored regions in p38 MAP kinase. (A) In complex with **3** (B) in complex with **4**. Regions s_{AP} and s_{BP} are depicted by the transparent blue volumes at r_c = full conditions (left) while s_{AP} and s_{BP} are depicted by red transparent volumes in r_c = bb conditions (right). X_{medoid} = 4 Å in both cases. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

and the cyano group is instead hydrogen-bonded to the two tyrosine phenolic hydroxyl groups (Fig. 6B). The monitored region in the p38 MAP kinase/**3** complex contains two water molecules that mediate hydrogen-bonding interactions between the ligand and the protein (Fig. 7A). Interestingly, the r_c = full and r_c = bb protocols lead to qualitatively different monitored regions. This is because in simulations of the complex with **3** under r_c = bb conditions, one of the two water molecules may sometime migrate to a third position, and then escape from the binding site. This occurred in *ca.* 5 ns in the first replicate, didn't occur in the second replicate, and occurred after 3 ns in the third replicate, but another water molecule returned after 20 ns to reproduce the original hydration state. This suggests a slow equilibrium between at least two hydration states. By contrast, the picture that emerges from simulation of **4** with the two restraining protocols is relatively consistent (Fig. 7B). In the case of EGFR kinase in complex with **5** (Fig. 8A), the monitored region contains a cluster of five water molecules in a tunnel that leads back to a solvent exposed surface of the protein. The monitored regions in the two restraining protocols are broadly similar, with the r_c = bb protocol leading to an enlarged monitored region owing to greater fluctuations in the positions of the water molecules. The cyano analogue **6** displaces a single water molecule as expected.

Binding energetics

Table 1 summarizes the components of the thermodynamic cycle depicted in Fig. 2, for varying restraint protocols and parameters that define the size of the monitored regions. The hydration free energies (rows 1–4) have been discussed previously. This data is completed with protein–ligand interaction energies (rows 5, 6), enabling computation of all the components (rows 7–12) of the thermodynamic cycle depicted in Fig. 2 for restraint protocols that feature heavy-atom restraints or limited restraints. Comparison of rows 7 and 8 indicate that while interaction energies are broadly consistent for scytalone dehydratase and EGFR kinase with the r_c = full or r_c = bb protocols, there is a significant variation in the case of p38 MAP kinase. Visualisation of the trajectories indicate that this occurs because **3** adopts a shifted binding mode owing to the occasional decreased water content of the monitored region, and protein side-chain rearrangements. Variations in protein–protein interaction energies are ignored in the present cycle and the result is unbalanced interaction energies between **3** and **4**. Row 11 and 12 lists the resulting binding site water displacement free energy for the three systems with the r_c = full or r_c = bb protocols. Both protocols indicate that the energetic cost for removing the water displaced by **6** in EGFR kinase is higher than for the displaced water molecules in scytalone dehydratase and p38 MAP kinase. However the free energy cost for displacing a water molecule from p38 MAP kinase is strongly influenced by restraints. This is because, as noted previously, in the r_c = bb protocol the water content of the monitored region exchanges slowly between states with one or two water molecules. Thus on average **4** displaces less than one water molecule under these conditions. The data in rows 11 and 12 can be compared with MC/FEP results from Michel *et al.*,[10] that reported MC/FEP water displacement free energies of

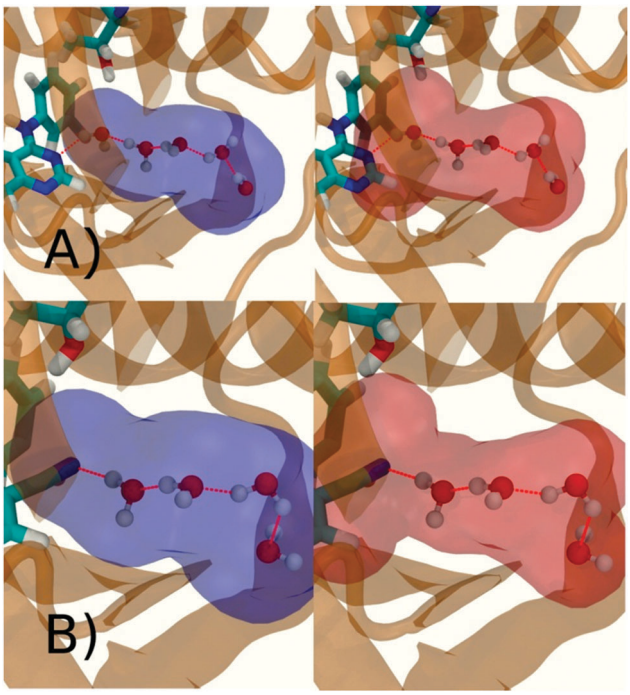

Fig. 8 Representation of GCT monitored regions in EGFR kinase (A) in complex with **5** (B) in complex with **6**. Regions s_{AP} and s_{BP} are depicted by the transparent blue volumes at r_c = full conditions (left) while s_{AP} and s_{BP} are depicted by red transparent volumes in r_c = bb (right) conditions. X_{medoid} = 4 Å in both cases. Relevant hydrogen-bonding interactions between protein residues, water molecules and ligands are depicted by red-dotted lines.

5.5 ± 0.2 kcal mol^{-1} (scytalone dehydratase), 4.2 ± 0.1 kcal mol^{-1} (p38 MAP kinase) and 6.9 ± 0.1 kcal mol^{-1} (EGFR kinase). Quantitative agreement is not expected as the forcefield and methods used differ, but qualitatively these figures are in closer agreement with those produced by the r_c = full protocol. Completing the cycle yields relative binding free energies (rows 13 and 14). The binding free energies are more precise for the full restraints protocol for scytalone dehydratase and p38 MAP kinase, but not EGFR kinase, presumably because of the larger number of water molecules in the monitored region of the latter protein. The variations of the computed relative binding energies are much greater than observed experimental data (row 15) or those obtained by previous MC/FEP calculations (row 16, albeit with a different forcefield).[31] GCT computed hydration free energies of small organic molecules have been shown previously to be highly correlated to TI computed hydration free energies. This suggests that the discrepancy here is likely due to the neglect of additional contributions such as changes in intramolecular energetics, or protein–ligand entropies, that would normally be included in a FEP/TI calculation. Others have also reported that the use of restraints tends to exaggerate the magnitude of the binding free energies of probe molecules to protein regions.[42] Nevertheless, the qualitative picture doesn't change, and the relative binding free energy for **5** → **6** is much less favourable than for **1** → **2** and **3** → **4**.

Entropic and enthalpic contributions to the energetics of binding site water displacement

The free energy change for water displacement was decomposed in enthalpic and entropic contribution. Fig. 9A indicates that in almost all cases the enthalpic component is unfavourable, whereas the entropic component is favourable regardless of the restraining protocol. The only exception is for **3** → **4** and r_c = bb, where the results are difficult to interpret since the number of water molecules displaced is on average less than one. The entropic component is relatively small and varies little across all systems, and variations in enthalpy changes dominate the overall thermodynamic signature. Fig. 9B breaks down

Table 1 Components of the thermodynamic cycle for evaluation of relative free energies of binding with the GCT approach. All figures are in kcal mol^{-1} and are quoted with one standard error of the mean

Protein	Scytalone dehydratase		p38 MAP kinase		EGFR kinase	
Ligand	**1**	**2**	**3**	**4**	**5**	**6**
$\Delta G^{s_{(A\|B)P}}_{w,(A\|B)P(r_c)}$, r_c = bb, X_{medoid} = 4 Å	−3.2 ± 1.2	0	−7.9 ± 1.5	−8.2 ± 0.4	−39.6 ± 2.9	−32.9 ± 3.1
$\Delta G^{s_{(A\|B)P}}_{w,(A\|B)P(r_c)}$, r_c = full, X_{medoid} = 4 Å	−4.2 ± 0.2	0	−12.70 ± 0.02	−6.2 ± 0.1	−36.4 ± 0.4	−27.9 ± 1.8
$\Delta G^{s_{(A\|B)}}_{w,(A\|B)P(r_l)}$, r_l = full, X_{vdw} = 6 Å	−21.2 ± 0.8	−19.6 ± 0.8	−46.6 ± 0.5	−43.9 ± 0.6	−11.8 ± 0.2	−19.1 ± 0.4
$\Delta G^{s_{(A\|B)}}_{w,(A\|B)P(r_l)}$, r_l = rot, X_{cubic} = 9 Å	−17.7 ± 0.8	−19.3 ± 1.5	−43.5 ± 1.3	−39.4 ± 0.8	−17.4 ± 1.5	−17.9 ± 0.6
ΔE(AP → BP, r_c), r_c = bb	−59.9 ± 0.2	−72.3 ± 0.2	−79.7 ± 0.2	−82.2 ± 3.1	−57.9 ± 0.3	−65.8 ± 0.4
ΔE(AP → BP, r_c), r_c = full	−61.4 ± 0.1	−75.0 ± 0.1	−68.4 ± 0.3	−83.7 ± 0.2	−54.7 ± 0.1	−62.0 ± 0.1
$\Delta\Delta E$(AP → BP, r_c), r_c = bb	−12.39 ± 0.04		−2.5 ± 3.0		−8.0 ± 0.5	
$\Delta\Delta E$(AP → BP, r_c), r_c = full	−13.6 ± 0.2		−15.3 ± 0.2		−7.2 ± 0.1	
$\Delta\Delta G_{hyd}$(A → B, s_A, s_B, r_l), r_l = full, X_{vdw} = 6 Å	−1.6 ± 1.7		−2.7 ± 0.2		7.3 ± 0.7	
$\Delta\Delta G_{hyd}$(A → B, s_A, s_B, r_l), r_l = rot, X_{cubic} = 9 Å	1.6 ± 1.5		−4.1 ± 1.2		0.5 ± 1.5	
$\Delta\Delta G_{hyd}$(AP → BP, s_{BP}, s_{AP}, r_c), r_c = full, X_{medoid} = 4 Å	4.4 ± 0.1		6.48 ± 0.04		8.5 ± 1.7	
$\Delta\Delta G_{hyd}$(AP → BP, s_{BP}, s_{AP}, r_c), r_c = bb, X_{medoid} = 4 Å	3.2 ± 1.2		−0.3 ± 1.4		6.7 ± 1.7	
$\Delta\Delta G_b$(A → B, s_{BP}, s_{AP}, s_A, s_B, r_c, r_l), r_c = full, X_{medoid} = 4 Å, r_l = full, X_{vdw} = 6 Å	−10.8 ± 1.7		−11.6 ± 0.4		8.6 ± 2.2	
$\Delta\Delta G_b$(A → B, s_{BP}, s_{AP}, s_A, s_B, r_c, r_l), r_c = bb, X_{medoid} = 4 Å, r_l = rot, X_{cubic} = 9 Å	−8.2 ± 2.9		−6.8 ± 2.6		−0.8 ± 2.2	
$\Delta\Delta G_b$ experimental	−2.0		−2.5		0.6	
$\Delta\Delta G_b$ MC/FEP study[31] (OPLS-AA/TIP4P)	−1.2 ± 0.2		−3.0 ± 0.3		1.4 ± 0.2	

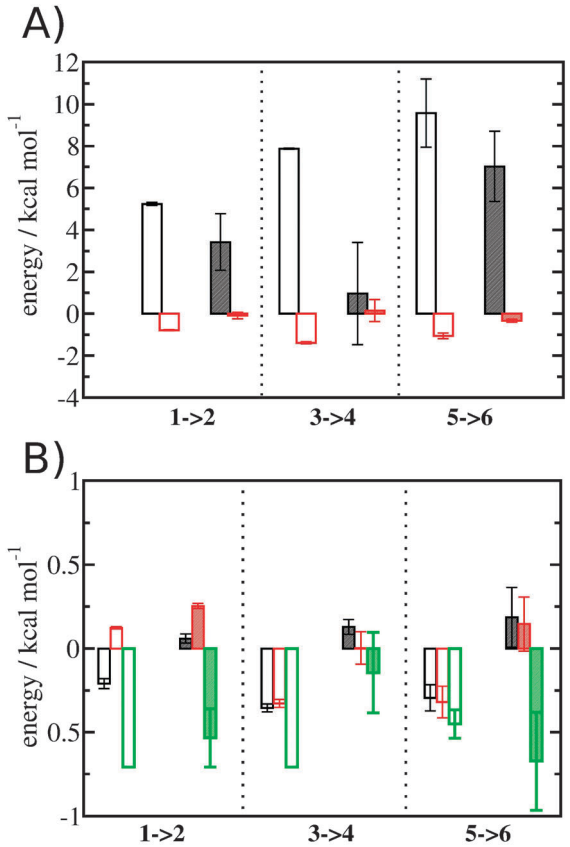

Fig. 9 Thermodynamic signature of the changes in the hydration energetics of the three protein–ligand complexes. (A) Enthalpy changes $\Delta\Delta H_{hyd}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), are shown as empty ($\boldsymbol{r}_c$ = full) or shaded ($\boldsymbol{r}_c$ = bb) black histograms. Entropy changes $-T\Delta\Delta S_{hyd}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), are shown as empty ($\boldsymbol{r}_c$ = full) or shaded ($\boldsymbol{r}_c$ = bb) red histograms. (B) Decomposition of the entropy changes in vibrational entropy $-T\Delta\Delta S_{vib}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (black), librational entropy $-T\Delta\Delta S_{lib}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (red) and orientational entropy $-T\Delta\Delta S_{ori}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) (green) components. Error bars represent the standard error of the mean from three replicates.

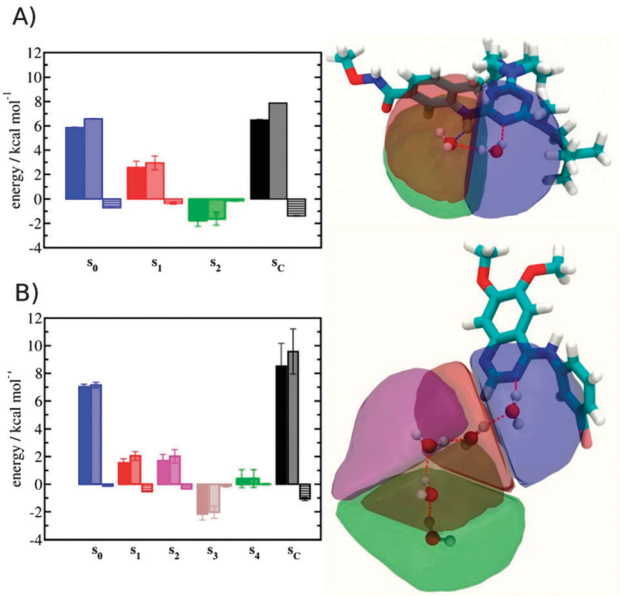

Fig. 10 Spatial decomposition of the changes in hydration energetics within the GCT monitored regions. The monitored region depicted in Fig. 7 and 8 was broken down into sub-regions for the $\boldsymbol{r}_c$ = full protocol simulations. For each sub-region, the relative hydration free energy $\Delta\Delta G_{hyd}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$), relative hydration enthalpy $\Delta\Delta H_{hyd}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) and relative hydration entropy $-T\Delta\Delta S_{hyd}$(AP → BP, $\boldsymbol{s}_{BP}$, $\boldsymbol{s}_{AP}$, $\boldsymbol{r}_c$) are depicted as bars (left). The contributions from the full region $\boldsymbol{s}_c$ are shown in black. The right panel depicts the localisation of each sub-region. (A) p38 MAP kinase. (B) EGFR kinase.

further the entropy changes into vibrational, librational and orientational components. The results indicate that displacing a water molecule may increase or decrease the vibrational and librational water entropy depending on the binding site and the simulation protocol, but the orientational entropy component dominates the overall entropy variations. This indicates that the favourable entropic contribution upon water displacement is due to the increased number of hydrogen-bonding orientations available to water in bulk.

Localisation of perturbations in water energetics

Additional insights into the binding process are gained by spatial decomposition of the hydration energetics of the monitored regions $\boldsymbol{s}_c$ into sub-regions. Fig. 10A shows that for p38 MAP kinase, the largest contribution arise from the volume of space $\boldsymbol{s}_0$ (blue) that was occupied by the water molecule displaced by **4**. The cyano group additionally perturbs the interactions of the neighbouring water molecule, shifting it from region $\boldsymbol{s}_1$ (red) towards $\boldsymbol{s}_2$ (green). The net effect almost cancels out and the energetic contributions from $\boldsymbol{s}_0$ are very similar to the full monitored region $\boldsymbol{s}_c$.

In EGFR kinase (Fig. 10B) the water volume displaced by the cyano group of **6** (blue) also accounts for the majority of the changes in hydration energetics. Additionally, the first hydration (red) and second (purple) hydration shells of the cyano group are destabilized, whereas the third (maroon) hydration shell is stabilised, and the fourth hydration shell (green) is unperturbed. Thus introduction of the cyano group has perturbed water properties up to 10 Å away. Here water network perturbations (all regions $\boldsymbol{s}_i$, $i > 0$) contribute approximately 1 additional kcal mol^{-1} to the changes in hydration energetics. Thus, that the **5** → **6** substitution is not energetically favourable is the result of: higher water displacement energetics (Fig. 10A $\boldsymbol{s}_0$ *versus* Fig. 10B $\boldsymbol{s}_0$), water network rearrangement penalties (Fig. 10B $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, $\boldsymbol{s}_3$), and weaker improvements in protein–ligand interaction energies (Table 1, row 8).

Discussion

The present study analysed in details the consequences of the use of different restraint protocols to control the allowed flexibility of protein and ligand atoms over the course of an MD simulation. Restraints are undesirable in the sense that they are artificial, and as the results have shown, can quantitatively and qualitatively affect the outcome of a GCT analysis. On the other hand, limited or lack of restraints, that should give

more accurate results, leads actually to poor reproducibility of computed quantities for simulations on a *ca.* 10 ns timescale. An important consideration of the present study was to explore the feasibility of using GCT for routine analyses in the context of structure-based drug design programs where computation is typically asked to inform evaluation of hundreds of candidate compounds on a time-scale of a few days. In this context, very long MD simulations are not practical. Overall the results suggest that for thermodynamic cycle analyses restraints should be used to probe specific protein–ligand conformational states. If different binding modes are to be evaluated, this is best done by separate analyses of different conformational states with r_c = full restraints, alternatively prohibitively long simulations may be needed to average over multiple binding modes, as evidenced for **3** with the protocol that enabled side-chain and ligand flexibility in p38 MAP kinase. If the expected binding modes are unknown, they could be explored prior analyses by means of unrestrained MD simulations. Additionally, care should be taken when selecting a representative conformation of the ligand for solution calculations, as evidenced by the discrepancy in computed relative hydration free energies for **5** and **6**.

Arguably, the appeal of GCT is in the additional information that it provides over, for instance, an alchemical relative hydration free energy calculation. The breakdown of hydration free energies into enthalpic and entropic components revealed that the variations in hydration energetics upon water displacement are dominated by enthalpy. A rationale for displacing water molecules from binding sites is the associated gain in entropy that should favour the process. However the data shown in Fig. 9 shows that this outcome, at least for the cases investigated here, may only be achieved if the relatively larger loss of enthalpy is counter-balanced by equally favourable additional protein–ligand interaction energies. In essence, harnessing entropy by water displacement requires carefully maintaining an energetically similar pattern of hydrogen-bonding interactions at the site of the displaced water molecule. The entropy gains are dominated by a favourable increase in orientational entropy and this is due to the lower average number of orientations that a water molecule may adopt in a binding site *versus* bulk conditions. Such observations have been reported for water in other binding sites,[27,43] and for a range of idealised host–guest cavities.[27] While it is possible to evaluate enthalpic and entropic contributions to free energies of binding of water molecules with FEP/TI this would require many more simulations at multiple temperatures,[44] and this route doesn't provide a breakdown of entropic contributions into physically insightful translational, rotational and orientational motions.

An important additional insight into the physical chemistry principles that underpin water-mediated protein–ligand interactions is provided by Fig. 10. In both p38 MAP kinase and EGFR kinase, most of the change in hydration free energy due to water displacement comes from the water molecule that was displaced by the cyano group of **4** and **6** respectively. However, further analysis of the neighboring solvent regions reveal that large but compensating variations in water energetics occurred. In the case of EGFR kinase, the perturbations in water properties propagate up to the third hydration shell of the cyano moiety, and these water network perturbations penalize additionally water displacement by approximately 1 kcal mol^{-1}. Investigation of other systems is desirable to establish the magnitude and frequency of water network perturbation effects in protein–ligand complexes.

Conclusions

The GCT methodology was developed to provide insights into the hydration thermodynamics of organic and biomolecules. Here it was applied for the first time to a set of protein–ligand complexes where congeneric ligand pairs displace a single water molecule from the binding site. It was shown that protocols that restrain the range of allowed motions of the protein and ligand may be more judicious in context where throughput and speed considerations are important, as it is for applications to structure-based drug design programs. More realistic models (*i.e.* fewer or no restraints) will require significantly longer simulations to achieve reasonable reproducibility. While hydration free energies can be predicted with a range of methodologies, the appeal of the GCT technique is that it provides insights into the contributions of enthalpy and entropy to the free energy changes, and that it enables a spatial decomposition of these components. This was used here to determine the spatial extent of the energetics perturbations in a water network upon modification of the chemical structure of a ligand. Further developments of the GCT formalism would be desirable to account for associated changes in protein and ligand entropy,[45] and to automatically assess the conformational dependence of hydration free energies. Overall the current GCT implementation appears well suited for clarifying the role of water in protein–ligand binding, and applications in combination with for instance alchemical free energy methods,[46] should be envisioned.

Acknowledgements

J.M. is supported by a University Research Fellowship from the Royal Society. This research was also supported by EPSRC through an award of a CASE studentship to G.G. and by a Royal Society research equipment grant (No. RG110450). The authors acknowledge Zhesi Zhu for useful preliminary results from an earlier study. The research leading to these results has also received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement No. 336289 to J.M.

Notes and references

1 W. L. Jorgensen, *Science*, 2004, **303**, 1813–1818.
2 J. Michel, *Phys. Chem. Chem. Phys.*, 2014, **16**, 4465–4477.
3 J. Michel, N. Foloppe and J. W. Essex, *Mol. Inf.*, 2010, **29**, 570–578.

4 A. Woodhead, H. Angove, M. Carr, G. Chessari, M. Congreve, J. Coyle, J. Cosme, B. Graham, P. Day, R. Downham, L. Fazal, R. Feltell, E. Figueroa, M. Frederickson, J. Lewis, R. McMenamin, C. Murray, A. O'Brien, L. Parra, S. Patel, T. Phillips, D. Rees, S. Rich, D.-M. Smith, G. Trewartha, M. Vinkovic, B. Williams and A. Woolford, *J. Med. Chem.*, 2010, **53**, 5956–5969.
5 M. Adler, D. D. Davey, G. B. Phillips, S. H. Kim, J. Jancarik, G. Rumennik, D. R. Light and M. Whitlow, *Biochemistry*, 2000, **39**, 12534–12542.
6 N. Huang and B. K. Shoichet, *J. Med. Chem.*, 2008, **51**, 4862–4865.
7 Z. Li and T. Lazaridis, *J. Phys. Chem. B*, 2004, **109**, 662–670.
8 D. Alvarez-Garcia and X. Barril, *J. Med. Chem.*, 2014, **57**, 8530–8539.
9 K. W. Lexa and H. A. Carlson, *J. Chem. Inf. Model.*, 2013, **53**, 391–402.
10 W. Yu, S. K. Lakkaraju, E. P. Raman and A. D. MacKerell Jr., *J. Comput.-Aided Mol. Des.*, 2014, **28**, 491–507.
11 J. Michel, J. Tirado-Rives and W. Jorgensen, *J. Phys. Chem. B*, 2009, **113**, 13337–13346.
12 M. S. Bodnarchuk, R. Viner, J. Michel and J. W. Essex, *J. Chem. Inf. Model.*, 2014, **54**, 1623–1633.
13 T. Lazaridis, *J. Phys. Chem. B*, 1998, **102**, 3531–3541.
14 T. Lazaridis, *J. Phys. Chem. B*, 1998, **102**, 3542–3550.
15 C. Nguyen, T. K. Young and M. Gilson, *J. Chem. Phys.*, 2012, **137**, 044101.
16 C. N. Nguyen, A. Cruz, M. K. Gilson and T. Kurtzman, *J. Chem. Theory Comput.*, 2014, **10**, 2769–2780.
17 T. Young, R. Abel, B. Kim, B. J. Berne and R. A. Friesner, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 808–813.
18 J. F. Truchon, A. Nicholl's, J. A. Grant, R. I. Iftimie, B. Roux and C. I. Bayly, *J. Comput. Chem.*, 2010, **31**, 811–824.
19 J.-F. Truchon, B. M. Pettitt and P. Labute, *J. Chem. Theory Comput.*, 2014, **10**, 934–941.
20 S. Zhou, L. T. Cheng, J. Dzubiella, B. Li and J. A. McCammon, *J. Chem. Theory Comput.*, 2014, **10**, 1454–1467.
21 M. A. Brodney, G. Barreiro, K. Ogilvie, E. Hajos-Korcsok, J. Murray, F. Vajdos, C. Ambroise, C. Christoffersen, K. Fisher, L. Lanyon, J. Liu, C. E. Nolan, J. M. Withka, K. A. Borzilleri, I. Efremov, C. E. Oborski, A. Varghese and B. T. O'Neill, *J. Med. Chem.*, 2012, **55**, 9224–9239.
22 S. J. Irudayam and R. H. Henchman, *Mol. Phys.*, 2011, **109**, 37–48.
23 S. Irudayam, R. Plumb and R. Henchman, *Faraday Discuss.*, 2010, **145**, 467–485.
24 R. Henchman, *J. Chem. Phys.*, 2007, **126**, 064504.
25 S. J. Irudayam and R. H. Henchman, *J. Phys.: Condens. Matter*, 2010, **22**, 284108.
26 G. Gerogiokas, G. Calabro, R. H. Henchman, M. W. Y. Southey, R. J. Law and J. Michel, *J. Chem. Theory Comput.*, 2014, **10**, 35–48.
27 J. Michel, R. H. Henchman, G. Gerogiokas, M. W. Y. Southey, M. P. Mazanetz and R. J. Law, *J. Chem. Theory Comput.*, 2014, **10**, 4055–4068.
28 J. M. Chen, S. L. Xu, Z. Wawrzak, G. S. Basarab and D. B. Jordan, *Biochemistry*, 1998, **37**, 17735–17744.
29 C. Liu, S. T. Wrobleski, J. Lin, G. Ahmed, A. Metzger, J. Wityak, K. M. Gillooly, D. J. Shuster, K. W. McIntyre, S. Pitt, D. R. Shen, R. F. Zhang, H. Zhang, A. M. Doweyko, D. Diller, I. Henderson, J. C. Barrish, J. H. Dodd, G. L. Schieven and K. Leftheris, *J. Med. Chem.*, 2005, **48**, 6261–6270.
30 A. Wissner, D. M. Berger, D. H. Boschelli, M. B. Floyd, L. M. Greenberger, B. C. Gruber, B. D. Johnson, N. Mamuya, R. Nilakantan, M. F. Reich, R. Shen, H. R. Tsou, E. Upeslacis, Y. F. Wang, B. Q. Wu, F. Ye and N. Zhang, *J. Med. Chem.*, 2000, **43**, 3244–3256.
31 J. Michel, J. Tirado-Rives and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2009, **131**, 15403–15411.
32 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
33 D. Seeliger and B. L. de Groot, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 417–422.
34 H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, 2004, **120**, 9665–9678.
35 J. Wang, R. Wolf, J. Caldwell, P. Kollman and D. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
36 A. Jakalian, D. Jack and C. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.
37 D. Case, T. A. Darden, T. E. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, M. Crowley, R. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin and P. Kollman, *Amber 11*, University of California, San Francisco, 2010.
38 H. Loeffler, C. J. Woods and J. Michel, *FESetup 1.0*, http://ccpforge.cse.rl.ac.uk/gf/project/ccpbiosim/.
39 (*a*) C. J. Woods and J. Michel, *Sire Molecular Simulation Framework, Revision 1786*, 2013, http://siremol.org/Sire/Home.html; (*b*) P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, *J. Chem. Theory Comput.*, 2013, **9**, 461–469.
40 I. Tironi, R. Sperb, P. Smith and W. van Gunsteren, *J. Chem. Phys.*, 1995, **102**, 5451–5459.
41 H. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384–2393.
42 D. Alvarez-Garcia and X. Barril, *J. Chem. Theory Comput.*, 2014, **10**, 2608–2614.
43 S. J. Irudayam and R. H. Henchman, *J. Phys. Chem. B*, 2009, **113**, 5871–5884.
44 L. R. Olano and S. W. Rick, *J. Am. Chem. Soc.*, 2004, **126**, 7991–8000.
45 U. Hensen, F. Gräter and R. H. Henchman, *J. Chem. Theory Comput.*, 2014, **10**, 4777–4781.
46 J. Michel and J. W. Essex, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 639–658.