Accuracy Assessment and Automation of Free Energy Calculations for Drug Design

ARTICLE *in* JOURNAL OF CHEMICAL INFORMATION AND MODELING · NOVEMBER 2013

Impact Factor: 3.74 · DOI: 10.1021/ci4004199 · Source: PubMed

2 AUTHORS, INCLUDING:

Thomas Fox
Boehringer Ingelheim
48 PUBLICATIONS **10,143** CITATIONS

SEE PROFILE

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

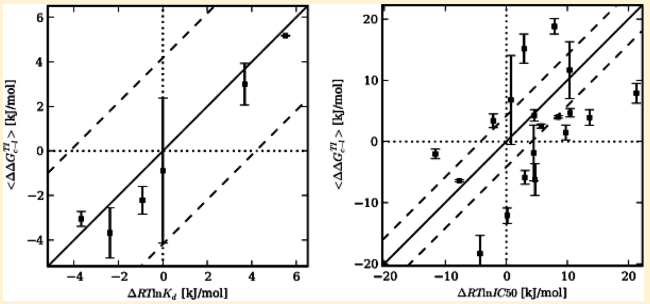
Accuracy Assessment and Automation of Free Energy Calculations for Drug Design

Clara D. Christ and Thomas Fox*

Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, 88397 Germany

Supporting Information

ABSTRACT: As the free energy of binding of a ligand to its target is one of the crucial optimization parameters in drug design, its accurate prediction is highly desirable. In the present study we have assessed the average accuracy of free energy calculations for a total of 92 ligands binding to five different targets. To make this study and future larger scale applications possible we automated the setup procedure. Starting from user defined binding modes, the procedure decides which ligands to connect via a perturbation based on maximum common substructure criteria and produces all necessary parameter files for free energy calculations in AMBER 11. For the systems investigated, errors due to insufficient sampling were found to be substantial in some cases whereas differences in estimators (thermodynamic integration (TI) versus multistate Bennett acceptance ratio (MBAR)) were found to be negligible. Analytical uncertainty estimates calculated from a single free energy calculation were found to be much smaller than the sample standard deviation obtained from two independent free energy calculations. Agreement with experiment was found to be system dependent ranging from excellent to mediocre (RMSE = [0.9, 8.2, 4.7, 5.7, 8.7] kJ/mol). When restricting analyses to free energy calculations with sample standard deviations below 1 kJ/mol agreement with experiment improved (RMSE = [0.8, 6.9, 1.8, 3.9, 5.6] kJ/mol).



1. INTRODUCTION

Binding affinity to the target is one crucial optimization parameter in drug design. Whereas other properties important for lead optimization are more readily accessible to computation, accurate predictions of free energies of binding remain challenging. If we could accurately predict binding affinity, we could fill in the current blank in the in silico characterization of compounds prior to synthesis. This would be particularly interesting in cases where the planned modifications require a new synthetic strategy as often the case for core modifications. With accurate binding affinity estimates at hand, the calculated property profile used to prioritize compounds for synthesis would not be lacking one of the most important optimization parameters anymore. Combining property prediction with virtual molecule generators[1] would allow focusing synthesis efforts onto the most promising molecules.

Recent years have brought a better understanding of the different challenges involved in the calculation of free energies of binding.[2−5] Blind prediction contests such as SAMPL3[6] have helped to get an idea of the currently achievable accuracy. Unfortunately, results were rather disappointing as shown, e.g., by the host−guest binding affinity prediction challenge where RMS errors varied from 5.9 to 190 kJ/mol.[7]

The purpose of this study was to assess the average accuracy of free energy calculations in drug design and to enable larger scale applications. We wanted to know how well-established methods work in our hands, in typical application scenarios, and without optimizing protocol or parameters to improve results for a specific system. To this end we automated the rather tedious setup of the calculations and determined the relative free energy of binding for a total of 92 ligands binding to five target proteins coming from a range of target classes relevant to drug design.

2. MATERIALS AND METHODS

2.1. Systems. The first data set consisted of six different ligands binding to the Mouse Major Urinary Protein-I (MUP-I) which are shown in Figure 1. The starting point for the simulations was the crystal structure of MUP-I in complex with 2-sec-butyl-4,5-dihydrothiazole (SBT) (1I06).[8] Amino acid protonation states corresponding to pH 7 were assigned using the Protonate 3D algorithm available in the molecular operating environment (MOE).[9] Binding modes for the other ligands were modeled based on SBT. Experimental data was taken from ref 10.

The second data set contained 32 substrate analog PDE5 inhibitors. Figure 2 shows the general scaffolds. Compound structures and experimental data were taken from ref 11. In the

Received: July 18, 2013
Published: November 20, 2013

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

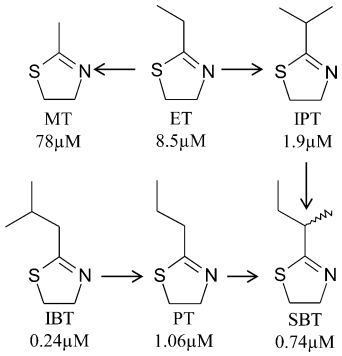


Figure 1. Structures and dissociation constants for the ligands of the MUP-I data set. Experimental data was taken from ref 10.

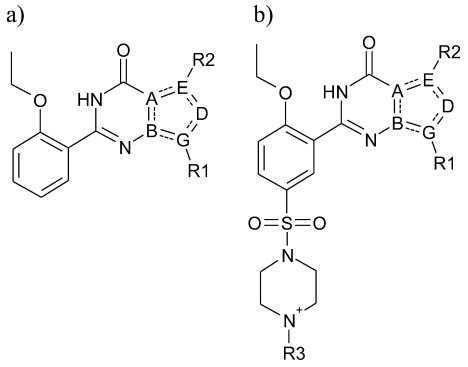


Figure 2. (a) Markush structure of the 15 2-ethoxyphenyl PDE5 inhibitors. (b) Markush structure of the 17 sulfonamide PDE5 inhibitors. Depictions analogous to ref 11.

following, compound numbering corresponds to the numbering in ref 11. Protein coordinates were taken from the crystal structure of PDE5A in complex with Sildenafil (1TBF).[12] Protonate 3D was used to assign protonation states corresponding to pH 7.[9] Ligand binding modes were modeled in MOE based on Sildenafil. All 17 sulfonamide PDE5 inhibitor perturbations were performed twice using the protonated form of the ligands (as shown in Figure 2b; denoted as "prot." in the following) as well as the neutral form.

Furthermore, three in-house data sets were used. The first in-house system contained 10 bromodomaine inhibitors. The second system consisted of 2 series of 9 and 14 serine protease inhibitors, respectively. The third system contained 20 phosphodiesterase inhibitors. Protonate 3D was used to assign amino acid protonation states corresponding to pH 7. Experimental data for these systems was obtained in-house.

2.2. Automated Simulation Setup Procedure. An automated procedure to setup all necessary files for free energy difference calculations in AMBER 11[13−15] was implemented using the OEChem toolkit.[16] The program takes a protein structure (as protein data bank file[17]) and a file with the corresponding ligand coordinates (as sd-file[18]) as input, i.e. the crucial definition of the binding modes is the responsibility of the user.

The first step consists of deciding which ligands should be perturbed into which other ligands. Manually one would most likely connect those ligands by a perturbation which differ the least. The underlying assumption being that smaller perturbations will on average converge faster than larger ones. While manually deciding which ligands should be connected by a perturbation is feasible for smaller sets it becomes very time-consuming for larger sets of molecules. The program therefore automates this choice. First, the maximum common substructure (MCS) of all pairs of ligands is computed. For each pair of ligands the number of common atoms (i.e., the number of atoms in the MCS) divided by the number of differing atoms plus one is calculated. This ratio is taken as a similarity score with higher numbers indicating more similar ligands and lower numbers indicating more dissimilar ligands. Using these scores, the maximum spanning tree which connects all ligands is calculated. Figure 3 depicts how the algorithm chooses which

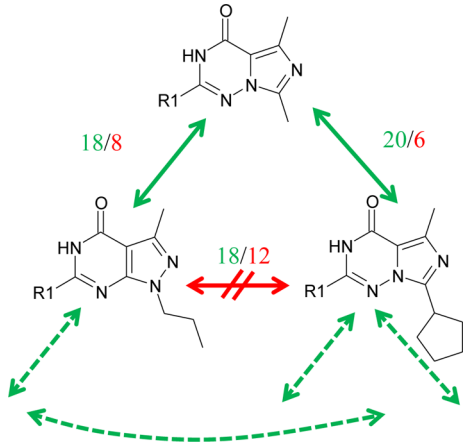


Figure 3. Pictorial representation of the construction of the maximum spanning tree of ligands based on the ratio of common (green numbers) and differing (red numbers) heavy atoms.

ligands to connect based on their similarity score. For each pair of ligands in the spanning tree, i.e. each perturbation to be performed, the necessary files for the relative free energy calculation are generated as described below and as depicted in Figure S1 of the Supporting Information. The arrows in Figure 1 show the tree that the algorithm determined for the MUP-I system.

Multiple maximum common substructures may be present for a ligand pair as illustrated in Figure 4. In order to avoid unnecessarily large perturbations, the algorithm chooses the MCS with the minimal total atom-to-atom distances.

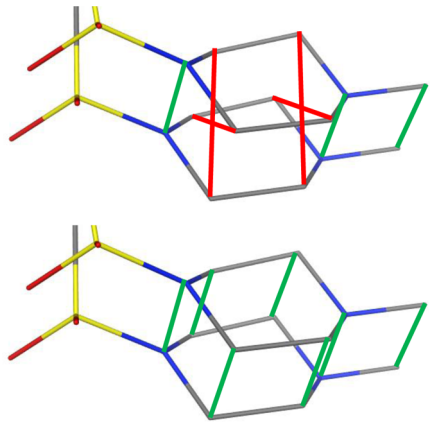


Figure 4. Pictorial representation of a ligand pair with multiple maximum common substructure matches due to symmetry. The upper MCS maps distant atoms (red lines); whereas, the lower MCS is the one with the lowest atom to atom distances (green lines). Ligand coordinates are shifted for clarity reasons.

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

Furthermore, atoms may be mapped on top of each other which are not close in space. For example, two stereoisomers would be considered identical by the MCS algorithm. Therefore, atoms further away in space than 0.5 Å as well as atoms differing in stereochemistry are removed from the MCS. For the two stereoisomers, therefore, the chiral center as well as the R-groups attached to it will be removed from the MCS.

AMBER 11 allows a simplified setup of dual topology[19] free energy calculations which does not need the specification of a perturbation topology. The user instead specifies two topology files. "Common atoms that are present in both states need to appear in the same order in both prmtop [topology] files and must have identical starting positions."[20,21] To ensure these requirements we adapt the atom coordinates and the atom and bond ordering. For each perturbation two sets of initial coordinates will be generated. For the first set, the coordinates of the MCS atoms of ligand 2 are set to the coordinates of the corresponding atoms in ligand 1. For the second set, the coordinates of the atoms in ligand 1 which are part of the MCS are set to the values of the corresponding atoms in ligand 2. In this way, two different starting configurations corresponding to the binding modes of ligand 1 and 2, respectively, are created for the subsequent simulations which both fulfill the requirement of having the same coordinates for common atoms. The two independent free energy simulations starting from these initial configurations allow a rough estimate of the sample standard deviation (see section 2.4). Ensuring the same atom order for common atoms turned out to be more challenging. When using `antechamber`[15,22,23] to generate parameters for the ligands, the so-called mainchain (the longest chain of atoms in the molecule) will determine the atom order. This mainchain may differ in the two ligands which shall be connected by a perturbation. It is therefore necessary to explicitly specify the mainchain. For performance reasons, we chose the longest of all shortest paths between all pairs of atoms in the MCS as the mainchain. Although this will in general not correspond to the longest chain within the molecule it is a suited chain of atoms for parameter generation. Ligand parameters are generated by first creating AM1BCC charges[24,25] using the program `antechamber`. AMBER PREP files are then generated with the program `prepgen`[15,23] using the antechamber output as well as the previously calculated mainchain. Additional force field parameters necessary for the molecule but not contained in the gaff force field[26] are generated using the program `parmchk`.[15,23] The setup script then calls `tleap`[15,23] to generate ligand and complex topologies as well as `sander`[15,21] input files and scripts to run the calculations. An example `tleap` input file can be found in listing S1 of Supporting Information. The AMBER force field `ff99SB`[27] and the general amber force field (gaff)[26] were used. To ensure identical water coordinates for both states, both ligands were loaded, combined and solvated in a box (truncated octahedron) of TIP3P[28] water with a distance of 12 Å between the solute and the edge of the periodic box. Then the second or the first ligand was removed and coordinates and parameters were saved for the first and second state, respectively. The same procedure was applied for the complex. Sodium and chloride ions were added to neutralize the system.[27,29,30]

2.3. Simulation Protocols. All simulations were performed using the AMBER 11[13−15] package for molecular dynamics simulation. Each relative free energy calculation consisted of 42 largely independent simulations. Figure 5 depicts how this large number arises: Simulations at 21 different lambda points have been performed both in the complex and in solution. Furthermore, this free energy calculation consisting of 42 largely independent simulations has been performed twice, starting from slightly different ligand conformations (see section 2.2).

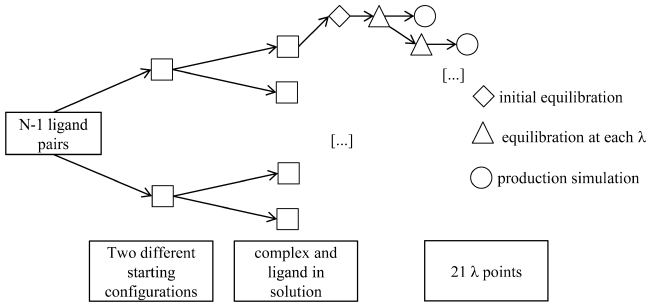


Figure 5. Depiction of the different simulations involved in the calculation of the relative binding affinities of *N* ligands.

All simulations were run as thermodynamic integration calculations using softcore potentials.[20,21] Charges and van der Waals parameters were perturbed simultaneously according to eqs 4.5−4.6 of ref 21 with softcore parameters $\alpha = 0.5$ and $\beta = 18$ Å. The softcore-atom specification `scmask` was generated by the setup procedure based on the atoms not contained in the MCS.

The initial equilibration consisted of 1000 steps of steepest decent minimization followed by 1000 steps of conjugate gradient minimization with all heavy atoms restrained (`restraint_wt` = 1000 Å^2 kcal/mol), followed by an equivalent minimization with only heavy atoms of the protein and/or the ligand restrained. Keeping these restraints the system was heated from 100 to 300 K within 50 ps at constant volume, then the box size was adjusted over 50 ps, followed by a cooling of the system to 100 K over 50 ps. In a series of nine minimizations equivalent to the one described above the restraints on protein and ligand heavy atoms were slowly released. The unrestrained system was then heated again from 100 to 300 K within 50 ps at constant volume followed by an adjustment of the box size over 50 ps. Further parameters were the following: all simulations were performed under periodic boundary conditions, the nonbonded cutoff was set to 9 Å, full electrostatic interactions were calculated using particle mesh Ewald[31,32] with AMBER 11 default parameters,[21] a time step of 2 fs was used, bonds involving hydrogens were constrained using the SHAKE[33] algorithm with a tolerance of 0.00001 Å. Please note that depending on the compiler used and compiler settings, sporadic SHAKE failures may occur in AMBER 11. In the present work this was largely prevented by using a `gcc4` compiled sander. This problem should have been resolved by AMBER 12 bugfix 8. Temperature was controlled using a Langevin thermostat[34] with a collision frequency of $\gamma = 3$ ps^{-1} and a different random seed for every run.[35,36] For constant pressure simulations, a weak-coupling algorithm[37] with a reference pressure of 1 bar, a compressibility of 44.6×10^{-6} bar^{-1}, and a relaxation time of 1 ps was used.

The initial equilibration was followed by a short 10 ps "equilibration at each λ" point (see Figure 5) under NPT conditions using the thermostat and barostat parameters described above. The subsequent longer "production simulation" consisted of 11 (MUP-I) and 16 (PDE5) runs of 100 ps each of which the first was skipped as further equilibration. This lead to a data collection phase of 1 ns (MUP-I) and 1.5 ns

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

(PDE5) at each of the 21 λ points. λ points were spaced with $\Delta\lambda$ = 0.05, however, with the extreme values shifted to λ = 0.005 and λ = 0.995, respectively, as AMBER 11 does not allow softcore thermodynamic integration simulations at λ = 0 or λ = 1 for technical reasons.[20,21] During the production simulation phase $\partial V/\partial\lambda$ values for thermodynamic integration[38] as well as potential energy values for subsequent multistate Bennett acceptance ratio MBAR[39] analysis were collected every 100 steps.

Summing the simulation time for both the ligand in solution and the ligand−protein complex in solution simulations over all λ points, the total simulation time for one estimate of the binding affinity difference was 47.12 (MUP-I) and 68.12 ns (PDE5), respectively. Note that for each perturbation, two such estimates starting from different ligand conformations were calculated (see section 2.2). The total simulation time for the MUP-I system (seven ligands including the two enantiomers of SBT, i.e. six perturbations) was therefore 565 ns and for the PDE5 system (47 perturbations) 6.4 μs.

For the first in-house system, simulation parameters were as for MUP-I. For in-house systems two and three simulation parameters differed slightly as follows as they were run in a project context prior to SHAKE problems being resolved (see above). Both systems were run without bond constraints and using a time step of 1 fs. The very short equilibration at each λ point (see Figure 5) was 5 ps long. It was followed by 16 runs of 50 ps each of which the first was skipped as further equilibration yielding a data collection phase of 0.75 ns per λ value.

2.4. Analysis. Free energy differences between two states for the ligand in solution (ΔG_l) and the protein−ligand complex (ΔG_c) were calculated using thermodynamic integration (TI) and the extension of the Bennett acceptance ratio method[40] to multiple states MBAR.[39]

For thermodynamic integration, the $\langle\partial V/\partial\lambda\rangle(\lambda)$ curves were linearly extrapolated from λ = 0.005 to 0 and from λ = 0.995 to 1.0.[20] The free energy difference (ΔG_l^{TI} or ΔG_c^{TI}) was then obtained by numerical integration using a simple trapezoidal rule. Subsequently, the difference in binding affinity was obtained as

$$\Delta\Delta G_{c-l}^{TI} = \Delta G_c^{TI} - \Delta G_l^{TI} \quad (1)$$

For each perturbation, two independent estimates for $\Delta\Delta G_{c-l}^{TI}$ were obtained and averaged:

$$\langle\Delta\Delta G_{c-l}^{TI}\rangle = \sum_{i=1}^{N} \frac{\Delta\Delta G_{c-l,i}^{TI}}{N} \quad (2)$$

(where N = 2). Note that averaging $\Delta\Delta G$ estimates is equivalent to averaging $\partial V/\partial\lambda$ values obtained from independent simulations as done in the independent-trajectories thermodynamic-integration (IT-TI) approach:[41]

$$\langle\Delta\Delta G_{c-l}^{TI}\rangle = \sum_{i=1}^{N} \frac{\Delta\Delta G_{c-l,i}^{TI}}{N} = \sum_{i=1}^{N} \frac{\Delta G_{i,c}}{N} - \sum_{i=1}^{N} \frac{\Delta G_{i,l}}{N} \quad (3)$$

$$= \sum_{i=1}^{N} \frac{\int_0^1 \langle\partial V_{i,c}(\lambda)/\partial\lambda\rangle_\lambda \, d\lambda}{N} - \sum_{i=1}^{N} \frac{\int_0^1 \langle\partial V_{i,l}(\lambda)/\partial\lambda\rangle_\lambda \, d\lambda}{N}$$

$$= \int_0^1 \frac{\sum_{i=1}^{N} \langle\partial V_{i,c}(\lambda)/\partial\lambda\rangle_\lambda}{N} d\lambda - \int_0^1 \frac{\sum_{i=1}^{N} \langle\partial V_{i,l}(\lambda)/\partial\lambda\rangle_\lambda}{N} d\lambda \quad (4)$$

where eq 3 corresponds to the approach taken in this study (with N = 2) and each of the terms in eq 4 correspond to the IT-TI equation (see eq 2 of ref 41). In the present approach, the two independent simulations differed slightly in their starting configurations; whereas in IT-TI independent trajectories were obtained by assigning different random velocities at simulation start up.

Uncertainties of the TI estimates were calculated as the sample standard deviation

$$\sigma_{\Delta\Delta G_{c-l}^{TI}} = \sqrt{\frac{\sum_{i=1}^{N} (\langle\Delta\Delta G_{c-l}^{TI}\rangle - \Delta\Delta G_{c-l,i}^{TI})^2}{N-1}} \quad (5)$$

where N = 2 and $\langle\Delta\Delta G_{c-l}^{TI}\rangle$ is the mean over the two estimates obtained starting from different initial configurations (see section 2.2). Furthermore, analytical uncertainty estimates were calculated for a limited set of perturbations for comparison. In order to obtain sensible estimates for the uncertainties, the collected data was subsampled as described in Appendix A of ref 39 to obtain a set of uncorrelated configurations. Analytical uncertainties $\delta_{\Delta G}$ were then calculated following Paliwal et al.[42,43] (see eq 4 and the Appendix in ref 42). Their approach of expressing the trapezoidal integration as a weighted sum of individual $\langle\partial V/\partial\lambda\rangle$ contributions allows a correct propagation of the variance obtained at the different λ points into ΔG. The obtained analytical uncertainty estimates for ΔG were propagated into $\Delta\Delta G$ as follows

$$\delta_{\Delta\Delta G_{c-l,ne}^{TI}} = \sqrt{\delta_{\Delta G_{c,ne}^{TI}}^2 + \delta_{\Delta G_{l,ne}^{TI}}^2} \quad (6)$$

where "ne" denotes not extrapolated and indicates a result obtained over the lambda range λ = 0.005−0.995.

Using the collected potential energy differences, MBAR estimates for the free energy differences in the range λ = 0.005−0.995 were calculated using eq 11 of ref 39. Analytical estimates of the associated uncertainties were calculated using eq 12 of the same reference. The complete MBAR analysis was performed using `pyMBAR`.[43] As AMBER currently does not allow softcore thermodynamic integration simulations at λ = 0 or 1 for technical reasons,[20,21] MBAR estimates could only be obtained in the range λ = 0.005−0.995. They can therefore not be compared to experiment and will only be compared to the TI estimates over the same range.

Postprocessing was performed in KNIME.[44] Pearson's product moment correlation coefficients including confidence intervals (denoted r and 95% CI in the following) were calculated using the R function "cor.test"[45] from within knime. Note that throughout the manuscript r rather than the

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

coefficient of determination r^2 is given. Root-mean-square errors (RMSE) and deviations (RMSD) were calculated as $\mathrm{RMSE(D)} = (\sum_{i=1}^{n}(x_i - y_i)^2/n)^{1/2}$.

3. RESULTS AND DISCUSSION

In the present work, we assess the accuracy of free energy calculations for drug design for a total of 92 ligands binding to 5 different target proteins. To make this study and future larger scale applications possible, an automated setup procedure was developed.

3.1. Dependence on Starting Conformation. To assess the dependence on the starting conformation each perturbation was performed twice starting from different ligand conformations corresponding to the binding mode model of ligand 1 and 2, respectively (see section 2.2). Figure 6a and b shows that the estimates for ΔG^{TI} obtained from run 1 and run 2 are perfectly correlated (r = 1.00) as they should be. [Here and in the following, if not denoted otherwise, analyses show data for all systems. For the PDE5 system only results for unprotonated ligands are shown. Furthermore results with > in experimental data are skipped. Detailed results for all systems can be found in Table 1 and in the Supporting Information.] For the perturbation of the free ligands in solution (ΔG_l^{TI}) a lower RMSE (0.6 kJ/mol) is observed than for the perturbation in the complex (ΔG_c^{TI}, RMSE = 3.1 kJ/mol). As the ΔG values are about an order of magnitude larger than their difference, the two independent estimates for $\Delta\Delta G_{c-l}^{\mathrm{TI}}$ correlate less well (r = 0.9) as can be seen in Figure 6c. The RMSE observed is of the same size as for ΔG_c^{TI}. The observed differences obtained when starting from differing ligand conformations confirm that incomplete sampling and therefore lack of convergence is a major error source in free energy calculations.[46] This lack of convergence is expected to increase when using starting conformations with larger differences as would be the case when using, e.g., different protein conformations. Our findings are in line with independent-trajectories thermodynamic-integration (IT-TI) results where it was found that "standard TI estimates from individual trajectories span a rather large free-energy estimate range".[41] The results obtained in the present study confirm that performing multiple independent simulations is crucial to detect and alleviate[47,48] sampling problems.

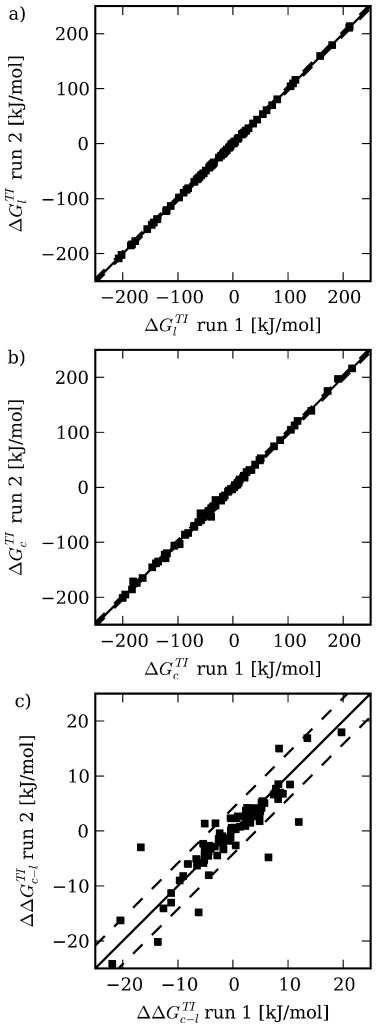


Figure 6. Dependence on starting conformation. (a) Free energy difference of the perturbation in solution (ΔG_l^{TI}) obtained from run 1 versus run 2 (N = 83, r = 1.0 (95% CI: 1.0|1.0), RMSE = 0.6 kJ/mol). (b) Free energy difference of the perturbation in the complex (ΔG_c^{TI}) obtained from run 1 versus run 2 (N = 83, r = 1.0 (95% CI: 1.0|1.0), RMSE = 3.1 kJ/mol). (c) Total free energy difference of the perturbation ($\Delta\Delta G_{c-l}^{\mathrm{TI}}$) from run 1 versus run 2 (N = 83, r = 0.9 (95% CI: 0.85|0.94), RMSE = 3.1 kJ/mol). The solid line shows $x = y$, dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

3.2. Extrapolation Error. Due to technical limitations in AMBER 11[20,21] $\langle \partial V/\partial\lambda \rangle$ curves needed to be extrapolated from λ = 0.005 to 0 and from λ = 0.995 to 1.0 prior to integration. To investigate the size of the error introduced by extrapolating, we compared $\Delta\Delta G_{c-l,\mathrm{ne}}^{\mathrm{TI}}$ obtained from nonextrapolated $\langle \partial V/\partial\lambda \rangle$ curves to the corresponding extrapolated values $\Delta\Delta G_{c-l}^{\mathrm{TI}}$. Figure 7 plots this comparison and shows that although 1% of the λ range is simply skipped for the nonextrapolated estimates, extrapolated and nonextrapolated values agree very well. The overall RMSD is only 0.3 kJ/mol and thus much smaller than the RMSE observed for the two independent runs (see section 3.1). For the presented data, the extrapolation error is therefore negligible.

3.3. Choice of Estimator. In the current study, differences between TI and MBAR estimates were found to be negligible. Figure 8 shows the very good agreement of TI and MBAR $\Delta\Delta G$ estimates (r = 0.995, RMSE = 0.7 kJ/mol). Note that the differences between the two independent runs were found to be much larger than the differences between the two estimators (see Figure 6c and the error bars in Figure 8). Although the differences between the two estimates we observed are very small, there is rationale to prefer the MBAR estimator over the TI estimator as it shows less bias when reducing the number of λ points.[42] Whether a reduction of λ points would be feasible without loss of accuracy should be addressed once MBAR estimates over the full lambda range (0−1) become available in AMBER (see section 2.3). The computational resources freed by using less λ points could then be used for better sampling at each λ point.

3.4. Uncertainty Estimation. Making large errors in binding affinity calculations would not be a problem in a drug design context if we could compute meaningful uncertainty estimates, i.e. uncertainty estimates which allowed us to distinguish reliable from unreliable predictions. Inaccuracies are expected to stem mainly from force field inaccuracies and insufficient sampling. As discussed above, we found that errors due to insufficient sampling can be substantial as indicated by a large sample standard deviation obtained from

Table 1. Calculated MBAR and TI Free Energy Differences for the Perturbation of the Ligands in Solution (ΔG_l) and in the Complex (ΔG_c), TI Estimates for the Difference in Binding Affinity ($\Delta\Delta G^{TI}_{c-l}$), and the Corresponding Experimental Affinity Differences ($\Delta RT \ln K_d$) for the MUP-I System[a]

system information			perturbation of ligand in solution			perturbation of complex in solution			difference in binding affinity			(all values in kJ/ mol)
			λ = 0.005− 0.995		λ = 0−1	λ = 0.005− 0.995		λ = 0−1	λ = 0−1			
compounds	starting conformation	number of perturbed atoms	$\Delta G^{MBAR}_{l,ne}$	$\Delta G^{TI}_{l,ne}$	ΔG^{TI}_{l}	$\Delta G^{MBAR}_{c,ne}$	$\Delta G^{TI}_{c,ne}$	ΔG^{TI}_{c}	$\Delta\Delta G^{TI}_{c-l}$	$\langle\Delta\Delta G^{TI}_{c-l}\rangle$	$\Delta RT \ln K_d$	|calc − expl
ET → IPT	IPT	7 → 10	−43.3 ± 0.1	−43.3	−43.8	−46.0 ± 0.1	−46.1	−46.6	−2.8	−3.1 ± 0.3	−3.7	0.6
	ET		−43.6 ± 0.1	−43.5	−44.0	−46.4 ± 0.1	−46.7	−47.3	−3.3			
ET → MT	MT	7 → 4	43.5 ± 0.1	43.3	43.8	48.5 ± 0.1	48.4	49.0	5.2	5.2 ± 0.0	5.5	0.3
	ET		43.4 ± 0.1	43.2	43.7	49.2 ± 0.1	48.4	48.9	5.2			
IBT → PT	PT	10 → 7	−34.2 ± 0.1	−35.1	−35.4	−30.4 ± 0.2	−31.5	−31.8	3.7	3.0 ± 0.9	3.7	0.7
	IBT		−35.0 ± 0.2	−35.0	−35.3	−32.9 ± 0.3	−32.7	−33.0	2.3			
IPT → SBT	SBT	6 → 9	23.8 ± 0.1	23.9	24.1	18.2 ± 0.2	19.5	19.6	−4.5	−3.7 ± 1.1	−2.4	1.3
	IPT		23.6 ± 0.1	23.7	23.9	19.9 ± 0.2	20.9	21.1	−2.9			
PT → SBT	SBT	10 → 13	−40.4 ± 0.2	−41.3	−41.8	−43.4 ± 0.3	−43.9	−44.5	−2.7	−2.2 ± 0.6	−0.9	1.3
	PT		−41.0 ± 0.2	−41.0	−41.4	−43.4 ± 0.3	−42.7	−43.2	−1.8			
SBT → SBT-R	SBT-R	13 → 13	−2.9 ± 0.1	−2.9	−2.9	−2.9 ± 0.4	−1.5	−1.5	1.4	−0.9 ± 3.3	0.0[b]	0.9
	SBT		−0.5 ± 0.1	−0.6	−0.6	−4.4 ± 0.3	−3.7	−3.8	−3.2			

[a]Experimental data taken from ref 10. The number of perturbed atoms $n_X \rightarrow n_Y$ indicates that n_X atoms are disappeared and n_Y appeared when perturbing compound $X \rightarrow Y$. Subscript ne indicates free energy differences in the range λ = 0.005−0.995, i.e., not extrapolated to 0 and 1 (see section 2.3). [b]According to ref 58, "both enantiomers of the pherhormone bind to MUP-I with approximately equal affinity".

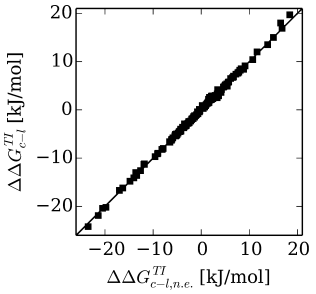


Figure 7. Comparison of $\Delta\Delta G^{TI}_{c-l,ne}$ obtained from nonextrapolated $\langle\partial V/\partial\lambda\rangle$ curves (λ = 0.005−0.995) with $\Delta\Delta G^{TI}_{c-l}$ obtained from extrapolated $\langle\partial V/\partial\lambda\rangle$ curves (λ = 0−1). N = 166, r = 1.00 (95% CI: 1.00|1.00), RMSD = 0.3 kJ/mol.

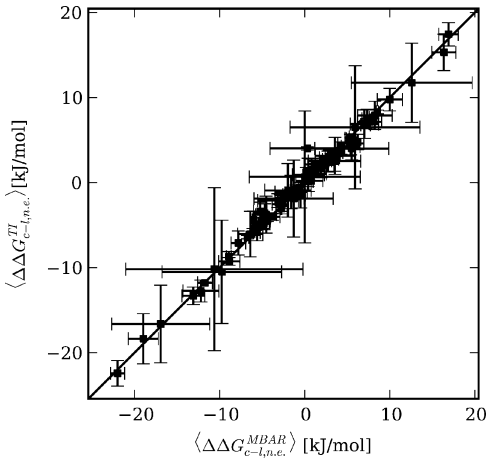


Figure 8. Choice of estimator. MBAR ($\langle\Delta\Delta G^{MBAR}_{c-l,ne}\rangle$) versus TI ($\langle\Delta\Delta G^{TI}_{c-l,ne}\rangle$). N = 83, r = 0.995 (95% CI: 0.992|0.997), RMSE = 0.7 kJ/mol. Error bars indicate the sample standard deviation over the two estimates starting from different initial conditions.

two independent free energy simulations starting from slightly different ligand conformations. Furthermore, we found that analytical uncertainty estimates obtained from a single free energy simulation underestimate the real sampling error. Figure 9 shows the sample standard deviation from the two

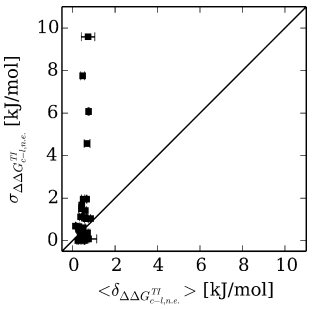


Figure 9. Comparison of two uncertainty estimates: Analytical ($\langle\delta\Delta\Delta G^{TI}_{c-l}\rangle$, obtained from uncorrelated configurations (see section 2.4)) versus sample standard deviation ($\sigma_{\Delta\Delta G^{TI}_{c-l,ne}}$) computed from two independent runs. Data obtained for the PDE5 system (unprotonated ligands only). Error bars indicate the sample standard deviation obtained from two runs. N = 31, r = 0.33 (95% CI: −0.03|0.61).

independent runs $\sigma_{\Delta\Delta G^{TI}_{c-l,ne}}$ versus the analytical uncertainty estimates $\delta_{\Delta\Delta G^{TI}_{c-l,ne}}$. The analytical uncertainty estimates are clearly much smaller than the observed $\sigma_{\Delta\Delta G^{TI}_{c-l,ne}}$. Hence, the observed precision in a single free energy simulation cannot be taken as a measure of the variability one would observe when running several free energy simulations with different starting configurations. The observed deviation of the different uncertainty estimates (sample standard deviation versus analytical) is a clear sign of suboptimal sampling of the system. Otherwise, no dependence on the starting conformation should be observed and different uncertainty estimates should be in agreement with each other as observed by Paliwal et al.[42] Here we have compared two different uncertainty estimates in the lambda range λ = 0.005−0.995 (ne) in order to allow a comparison. When comparing to experiment in the following,

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

we will switch to the whole lambda range, i.e. $\sigma_{\Delta\Delta G_{c-l}^{TI}}$. Please note that $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ and $\sigma_{\Delta\Delta G_{c-l,ne}^{TI}}$ are very similar: For the data set shown in Figure 9 their root-mean-square deviation is smaller than 0.1 kJ/mol.

Although $\sigma_{\Delta\Delta G_{c-l,ne}^{TI}}$ estimates are substantially larger than the analytical uncertainty estimates $\delta_{\Delta\Delta G_{c-l,ne}^{TI}}$, they underestimate the real error. Figure 10 shows the deviation of the predicted and

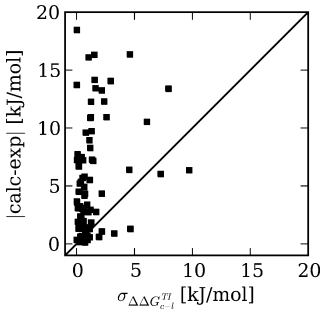


Figure 10. Sample standard deviation (of TI estimate, $\sigma_{\Delta\Delta G_{c-l}^{TI}}$) versus deviation of TI estimate from experiment (|calc − exp|), N = 83, r = 0.28 (95% CI: 0.07|0.47).

the experimental $\Delta\Delta G$ versus $\sigma_{\Delta\Delta G_{c-l}^{TI}}$. On the one hand $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ is expected to be smaller as it does not capture any force field inaccuracies. On the other hand two simulations are not enough to get a reliable estimate of $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ for the given simulation time. Yet, a small $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ is a necessary (not sufficient!) condition for convergence. In summary, large $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ indicate that agreement with experiment cannot be expected. Small $\sigma_{\Delta\Delta G_{c-l}^{TI}}$, however, cannot be taken as a guarantee that calculated and experimental free energy differences will agree within $\sigma_{\Delta\Delta G_{c-l}^{TI}}$.

3.5. Dependence on the Number of Perturbed Atoms.

In our setup procedure, we chose perturbations in such a way that similar ligands were connected. The rationale being that smaller perturbations should converge faster than larger ones. Figure 11 shows $\sigma_{\Delta\Delta G_{c-l}^{TI}}$ and the real error (i.e., deviation from experiment) as a function of the number of perturbed atoms.

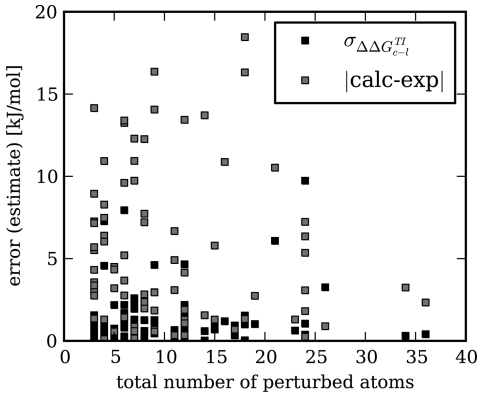


Figure 11. Dependence on the number of perturbed atoms: Sample standard deviation of TI estimate ($\sigma_{\Delta\Delta G_{c-l}^{TI}}$, black squares, N = 83, r = 0.03 (95% CI: −0.19|0.24)) and deviation from experiment (|calc − exp| = $|\langle\Delta\Delta G_{c-l}^{TI}\rangle - RT \ln IC_{50}|$, gray squares, N = 83, r = −0.04 (95% CI: −0.26|0.17)) versus total number of perturbed atoms.

Interestingly, no correlation between those quantities can be deduced from the data. This lack of correlation indicates that convergence is dominated by insufficient sampling of uncorrelated configurations rather than by the size of the perturbation.

Expecting smaller perturbations to converge faster than larger ones assumes that increasing the size of the perturbation decreases the overlap of relevant configurational space of the two systems. Although this might often be a reasonable approximation, it need not always be true. Two ligands might differ in just one core atom which however induces a completely different water and/or protein structure around the ligand turning a seemingly simple perturbation rather hard. Or two ligands might both have different bulky groups of similar nature, leading to a large number of perturbed atoms but possibly a rather simple perturbation. However, as the overlap of relevant configurational space is hard to access a priori, relying on minimizing the number of perturbed atoms still seems a reasonable approach.

3.6. Dependence on the Absolute Size of ΔG.

Binding affinity estimates are traditionally obtained by subtracting the free energy changes in the complex and in solution. This implies the subtraction of two large numbers of similar magnitude to obtain a difference which is about an order of magnitude smaller (see also Figure 6). Recent approaches prevent this subtraction which is prone to be imprecise by assessing $\Delta\Delta G$ directly.[49] In that work "a single-simulation alternative to the thermodynamic cycle that is standardly used" is presented which "relies on the coupled simulation of two systems".[49] In the present study however, $\Delta\Delta G$s were obtained in the traditional way from ΔG_c and ΔG_l. We wanted to investigate whether the error in the estimates increases with increasing $|\Delta G|$, as is sometimes stated. Figure 12 shows $\sigma_{\Delta\Delta G_{c-l}^{TI}}$

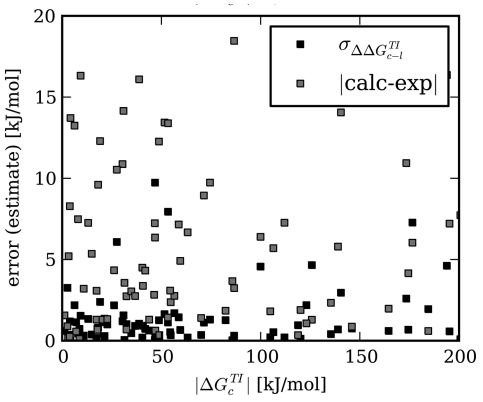


Figure 12. Dependence on the absolute size of ΔG: Sample standard deviation of TI estimate ($\sigma_{\Delta\Delta G_{c-l}^{TI}}$, black squares, N = 83, r = 0.07 (95% CI: −0.14|0.29)) and deviation from experiment (|calc − exp| = $|\langle\Delta\Delta G_{c-l}^{TI}\rangle - RT \ln IC_{50}|$, gray squares, N = 83, r = 0.11 (95% CI: −0.1|0.32)) versus absolute size of free energy change in the complex ($|\Delta G_c^{TI}|$). Analysis of the perturbation of the ligands in solution yields similar results.

and the deviation from experiment as a function of the absolute size of the single free energy difference in the complex ($|\Delta G_c^{TI}|$). Interestingly, no increase of the error with increasing $|\Delta G_c^{TI}|$ is observed for the investigated data set. The problem of subtracting two large ΔGs to obtain $\Delta\Delta G$ may actually not be as problematic after all.

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

3.7. Calculated versus Experimental Binding Affinities. *MUP-I.* For the system of seven ligands (including both enantiomers of SBT) binding to MUP-I excellent agreement with experiment was obtained (RMSE = 0.9 kJ/mol). Figure 13

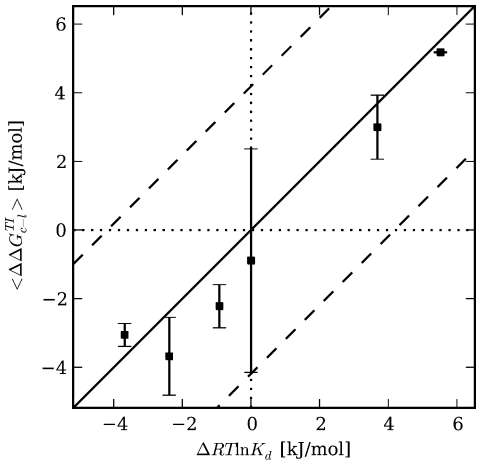


Figure 13. Calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ versus experimental ΔRT ln K_d for MUP-I. $N = 6$, $r = 0.98$ (95% CI: 0.82|1.0), RMSE = 0.9 kJ/mol. Error bars indicate the sample standard deviation $\sigma_{\Delta\Delta G^{TI}_{c-l}}$ obtained from two independent runs. The solid line shows $x = y$, dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

shows calculated TI versus experimental affinity differences, and Table 1 contains the detailed calculation results. When excluding calculations with higher uncertainty estimates ($\sigma_{\Delta\Delta G^{TI}_{c-l}} > 1$ kJ/mol), the RMSE is 0.8 kJ/mol.

For this rather small system with its simple perturbations quantitative predictions seem feasible. Note that the obtained results are far more accurate than the WaterMap plus cavity contribution and MM/GBSA results presented by Wang et al.[50] Results for four of the six perturbations can be derived from their data. The RMSE is 6.8 kJ/mol for their WaterMap plus cavity contribution approach and 5.7 kJ/mol for the MM/GBSA results they present compared to 1.0 kJ/mol for the TI calculations presented here.

PDE5. The PDE5 system unites two challenges often encountered in applications of free energy calculations: the treatment of tautomers and protomers. Whereas we do not tackle the former in the present study [indeed, ab initio calculations show that the 1-NH-Lactam is the prevalent tautomer in solution (data not shown), and it is this tautomer that forms an important hydrogen bond to an active site glutamine[12]], we did investigate the influence of the amine protonation state on the 17 sulfonamide PDE5 inhibitors which are most likely protonated. Ideally, one would like to calculate free energy differences including net charge changes; however, elaborate corrections are necessary to make these calculations independent of the electrostatic scheme used.[51,52] The form of these corrections for the ligand binding case have yet to be established.

Figure 14a shows the calculated free energy differences for the protonated sulfonamides versus the results obtained for the neutral sulfonamides. The differences between the two calculations are in the same range as the RMSE of run 1 versus run 2 for the whole data set (see Figure 6). Figure 14b and c compare the calculated differences for the protonated and neutral sulfonamide ligands, respectively, to experiment. The

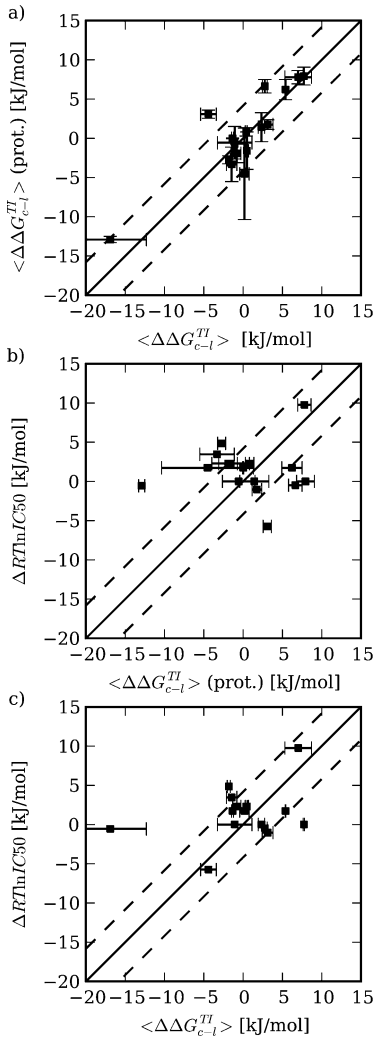


Figure 14. Comparison of results for protonated and corresponding neutral PDE5 inhibitors ($N = 16$). (a) $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ (prot.) versus $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ (neutral ligands) ($r = 0.87$ (95% CI: 0.65|0.95), RMSE = 2.8 kJ/mol). The middle and bottom plot show $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ versus experiment (ΔRT ln IC_{50}) for protonated (b), $r = 0.1$ (95% CI: −0.4|0.6), RMSE = 5.9 kJ/mol) and neutral (c), $r = 0.3$ (95% CI: −0.2|0.7), RMSE = 5.5 kJ/mol) ligands, respectively. Error bars indicate $\sigma_{\Delta\Delta G^{TI}_{c-l}}$. The solid line shows $x = y$, and dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

calculated affinity differences for the neutral ligands differ somewhat less than those calculated for the protonated ligands. As the ligands are most likely protonated under experimental conditions, one might expect to obtain better agreement for the protonated ligands. However, perturbations involving charged species are expected to have worse convergence properties than neutral ones due to strong electrostatic interactions in the system. This might explain why the model which should be closer to reality (the protonated ligands) does not yield superior results in this case.

The excellent correlation of calculated and experimental binding affinity differences achieved for the MUP-I system could not be obtained for the PDE5 system. Figure 15 relates calculated and experimental free energy differences. The detailed calculation results can be found in the Supporting Information. The data is scattered and no correlation can be observed. However, the overall RMSE of 8.2 kJ/mol (i.e., less

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

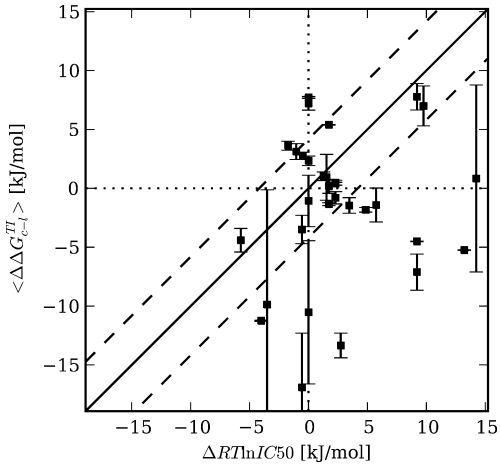


Figure 15. Calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ versus experimental ΔRT ln IC_{50} for PDE5 (unprotonated ligands only). $N = 31$, $r = 0.16$ (95% CI: −0.21| 0.49), RMSE = 8.2 kJ/mol. Error bars indicate $\sigma_{\Delta\Delta G^{TI}_{c-l}}$. The solid line shows $x = y$, and dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

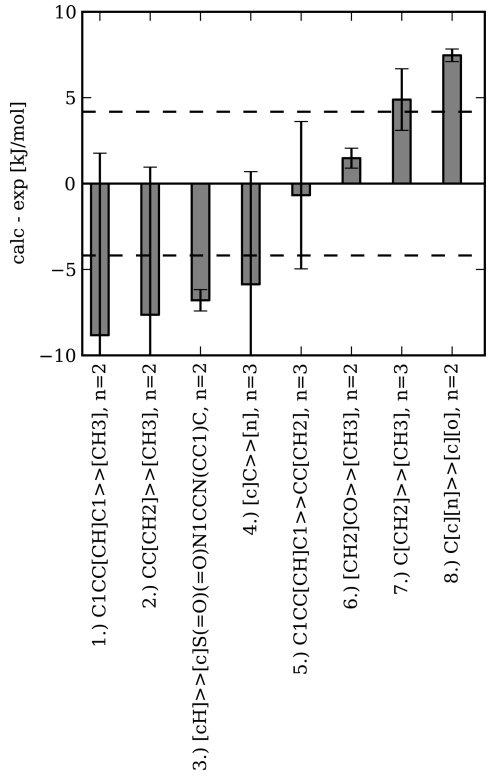


Figure 16. Difference between calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ and experimental ΔRT ln IC_{50} binned by perturbation type ($n \geq 2$) for the PDE5 system (unprotonated ligands only). Error bars indicate the standard deviation $\sigma_{calc-exp}$. Dashed lines show $y = \pm 4.184$ kJ/mol = ± 1 kcal/mol.

than 2 kcal/mol) is not particularly high compared to the findings of others.[7−53] Unfortunately, the obtained accuracy for this system is not good enough for these calculation to be useful in a drug design context. Expressed more positively, 52% of the free energy differences are within 4.184 kJ/mol (i.e., 1 kcal/mol) of experiment. When restricting the analysis to perturbations for which uncertainty estimates were low ($\sigma_{\Delta\Delta G^{TI}_{c-l}} < 1$ kJ/mol), this ratio improves slightly (to 56%) and the RMSE is reduced to 6.9 kJ/mol. This still rather high RMSE indicates that there are several perturbations which are precise but deviate strongly from experiment. One such perturbation is 11 → 13 which corresponds to the mutation of a propyl-group to a methyl-group in position R1 (see Figure 2, scaffold a). The calculated free energy difference is $\langle\Delta\Delta G^{TI}_{c-l}\rangle = -4.5 \pm 0.0$ kJ/mol (see SI Table S1) whereas the experimental value is ΔRT ln $IC_{50} = 9.2$. Interestingly, this same perturbation on scaffold b (25 → 26) matches the experimental value ($\langle\Delta\Delta G^{TI}_{c-l}\rangle = 0.2 \pm 0.6$ kJ/mol and ΔRT ln $IC_{50} = 1.7$). Figure S2 in the Supporting Information shows $\langle\partial V(\lambda)/\partial\lambda\rangle$ versus simulation time. The $\langle\partial V(\lambda)/\partial\lambda\rangle$ curve of the complex simulation is less smooth than the one of the ligand in solution simulation. This is not surprising and is also reflected in the analytical uncertainty estimate $\delta_{\Delta\Delta G^{TI}_{c-l,ne}}$ which is 0.5 kJ/mol for the complex and 0.2 kJ/mol for the ligand simulations. SI Figure S3 shows the free energy estimate $\Delta\Delta G^{TI}_{c-l}$ versus time. After 600 ps, the sample standard deviation $\sigma_{\Delta\Delta G^{TI}_{c-l}}$ falls below 1 kJ/mol which is of the same size as the variation observed in measured experimental RT ln IC_{50} over time.[54] Futhermore, MBAR and TI estimates for $\Delta G_{[c,l],ne}$ agree and ΔG contributions due to extrapolation are negligible (see SI Table S1). In summary, these observations show that there is no reason to question this particular $\Delta\Delta G^{TI}_{c-l}$ estimate. Further examinations, which are beyond the scope of this work, would be necessary to clarify the observed discrepancy.

Separating errors due to sampling and force-field parameters is challenging. Here, we assess the sampling error by performing each calculation twice and calculating the sample standard deviation ($\sigma_{\Delta\Delta G^{TI}_{c-l}}$; see eq 5). To assess errors in the force field parameters, Figure 16 shows the deviation of calculated and experimental binding affinity differences by perturbation type. Perturbations 1 and 2 correspond to mutating a cyclopentyl-ring to methyl and propyl to methyl (see above), respectively. These perturbations are on average not particularly accurate but the mean deviation is not precise enough to allow for any conclusion about force field inaccuracies. Perturbation 3 corresponds to the introduction of the large sulfonamide group (see Figure 2). Although the mean deviation from experiment does not scatter, one of the calculated values is not very precise (see perturbation 14 → 35 in SI Table S1). Perturbation 4 of a pyrazole into a triazole is on average not very accurate but again the mean is not precise enough to draw any conclusions. Mutating a cyclopentyl-group into an propyl-group works well on average (perturbation 5) as does perturbation 6 which modifies the sulfonamide substituent. The ethyl to methyl perturbation 7 in R2 also works reasonably well on average. Perturbation 8 which mutates an imidazole into an isoxazole is interesting as the deviation from experiment is rather large and does not scatter. This might hint to a suboptimal isoxazole parametrization.

In-House System I. For the first in-house system, all but one of the calculated binding affinity differences agree well with experiment (see Figure 17). The accuracy obtained (RMSE = 4.7) for this system would be sufficient for TI calculations to be useful in a drug design context. However, as can be seen from the one outlier, a couple of nicely matching calculated affinity differences do by no means indicate that subsequent calculations will also be accurate. When excluding calculations with higher uncertainty estimates ($\sigma_{\Delta\Delta G^{TI}_{c-l}} > 1$ kJ/mol), the RMSE drops to 1.8 kJ/mol.

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

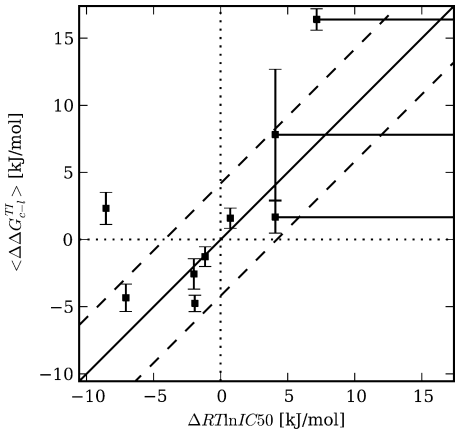


Figure 17. Calculated $\langle \Delta\Delta G^{TI}_{c-l} \rangle$ versus experimental ΔRT ln IC_{50} for in-house system I. $N = 6$ (skipping values with > in experimental data), $r = -0.043$ (95% CI: −0.83|0.8), RMSE = 4.7 kJ/mol. Error bars indicate $\sigma_{\Delta\Delta G^{TI}_{c-l}}$. The solid line shows $x = y$, and dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

The different perturbations involved in this system are shown in Figure 18. Perturbations 1−5 work nicely and correspond to

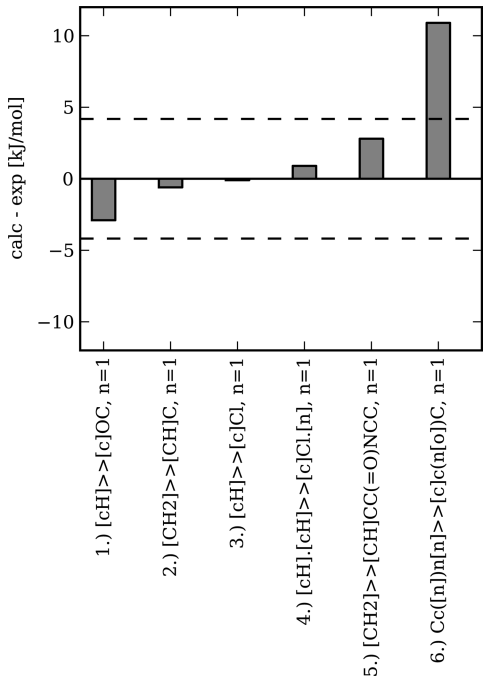


Figure 18. Difference between calculated $\langle \Delta\Delta G^{TI}_{c-l} \rangle$ and experimental ΔRT ln IC_{50} by perturbation type for in-house system one (skipping perturbations with > in experimental data). Dashed lines show $y = \pm$ 4.184 kJ/mol = ± 1 kcal/mol.

the introduction of a methoxy group, a methyl group, a chloro-substituent, a chloro-substituent plus a nitrogen in an aromatic ring, and an amide substituent, respectively. Perturbation 5 mutates a triazole into an isoxazole and is calculated to be unfavorable whereas experimentally it is said to be very favorable. As also in this system (as in PDE5) the perturbation involving isoxazole deviates strongly from experiment, the suspicion that the isoxazole parametrization is suboptimal is substantiated.

In-House System II. Figure 19 compares calculated versus experimental affinity differences for in-house system II. With an

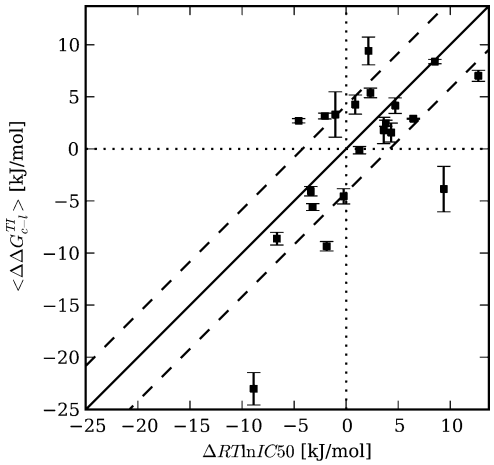


Figure 19. Calculated $\langle \Delta\Delta G^{TI}_{c-l} \rangle$ versus experimental ΔRT ln IC_{50} for in-house system II. $N = 21$, $r = 0.65$ (95% CI: 0.3|0.84), RMSE = 5.7 kJ/mol. Error bars indicate $\sigma_{\Delta\Delta G^{TI}_{c-l}}$. The solid line shows $x = y$, and dashed lines show $y = x \pm 4.184$ kJ/mol = $x \pm 1$ kcal/mol.

RMSE of 5.7 kJ/mol and correctly predicted signs for most perturbations, the accuracy obtained would be good enough to make these calculations useful in a drug design context. Of all calculated values, 57% are within 4.184 kJ/mol of experiment. When excluding calculations with higher uncertainty estimates ($\sigma_{\Delta\Delta G^{TI}_{c-l}} > 1$ kJ/mol), this ratio improves to 69% and the RMSE reduces to 3.9 kJ/mol.

The deviation from experiment binned by the perturbations involved is shown in Figure 20. The perturbations correspond to modification of aromatic substituents (1) hydrogen (H) to fluoro, (2) H to cyano, (3) removal of a chloro-substituent and

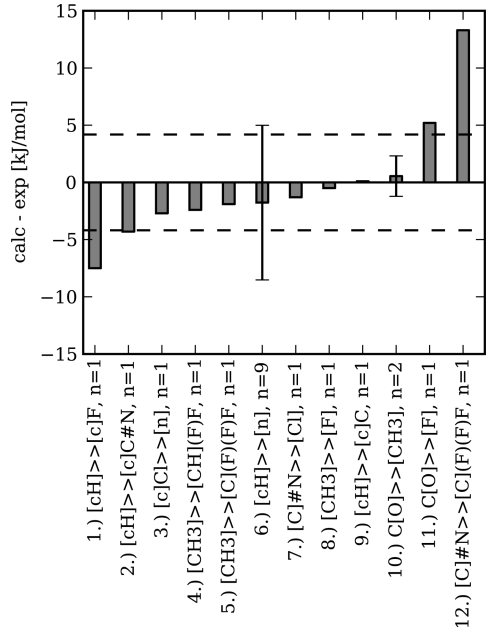


Figure 20. Difference between calculated $\langle \Delta\Delta G^{TI}_{c-l} \rangle$ and experimental ΔRT ln IC_{50} binned by perturbation type for in-house system II. Error bars indicate the standard deviation $\sigma_{calc-exp}$. Dashed lines show $y = \pm$ 4.184 kJ/mol = ± 1 kcal/mol.

dx.doi.org/10.1021/ci4004199 | J. Chem. Inf. Model. 2014, 54, 108−120

introduction of a ring nitrogen, (4) methyl to difluoromethyl, (5) methyl to trifluoromethyl, (6) introduction of a ring nitrogen, (7) cyano to chloro, (8) methyl to fluoro, (9) H to methyl, (10) methoxy to methyl, (11) methoxy to fluoro, (12) cyano to trifluoromethyl. Some of the perturbation involving cyano- and fluoro-compounds show larger deviation from experiment (1, 11, 12); others, however, do not (7, 8). For the introduction of a ring nitrogen (perturbation 6), no systematic deviation from experiment is observed but accuracy of the calculations varies.

In-House System III. Figure 21 shows experimental and calculated free energy differences for the third in-house system.

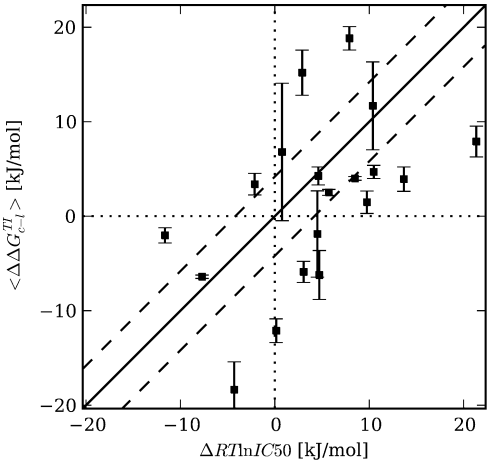


Figure 21. Calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ versus experimental ΔRT ln IC_{50} for in-house system III. $N = 19$, $r = 0.5$ (95% CI: 0.055|0.78), RMSE = 8.7 kJ/mol. Error bars indicate $\sigma_{\Delta\Delta G^{TI}_{c-l}}$. The solid line shows $x = y$, and dashed lines show $y = x \pm 4.184$ kJ/mol $= x \pm 1$ kcal/mol.

With an RMSE of 8.7 kJ/mol, agreement with experiment is suboptimal. Only 21% of all calculated free energy differences lie within 4.184 kJ/mol of the experimental value. Restricting the analysis to perturbations with low uncertainty estimates ($\sigma_{\Delta\Delta G^{TI}_{c-l}} < 1$ kJ/mol), the ratio improves only to 40% and the RMSE reduces to 5.6 kJ/mol with only 5 of 19 perturbations fulfilling the criterion. This low number indicates that simulation time for this system was not sufficient to reach that level of precision. Monitoring $\sigma_{\Delta\Delta G^{TI}_{c-l}}$ as a function of simulation time could be used to prolong production simulation time until a given level of precision is reached. Yet, at least the sign of most (74%) perturbations, i.e. whether a change will be favorable or not, is correctly predicted. However, whether calculations with this level of accuracy might still be useful in a drug design context is questionable.

When looking at the perturbations involved (Figure 22), it becomes clear that many of the perturbations involve sulfur and bromine. These perturbations are certainly harder to correctly describe with classical force fields than, e.g., the simple changes in the alkyl group of the MUP-I ligands.

For the comparisons above we have used the calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ values and compared them to the corresponding experimental binding affinity differences. Due to the construction of a spanning tree all $N(N-1)/2$ differences in free energy of binding of N ligands can be calculated from the $N-1$ calculated $\Delta\Delta G$s. However, it should be kept in mind that when calculating a binding affinity difference from several $\Delta\Delta G$s along the spanning tree, errors will add up.

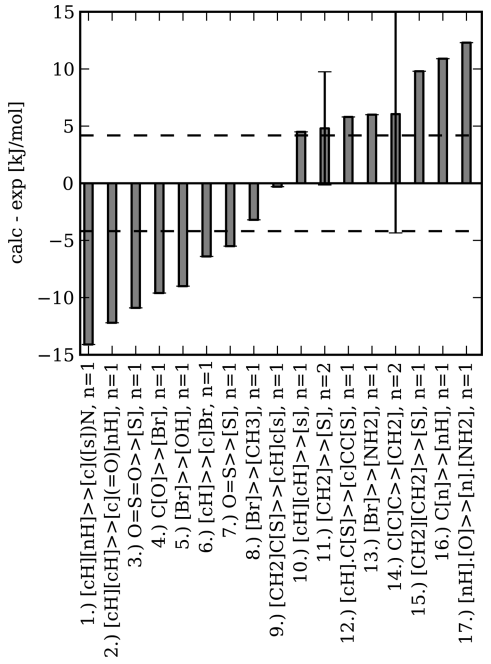


Figure 22. Difference between calculated $\langle\Delta\Delta G^{TI}_{c-l}\rangle$ and experimental ΔRT ln IC_{50} binned by perturbation type for in-house system III. Error bars indicate the standard deviation $\sigma_{calc-exp}$. Dashed lines show $y = \pm 4.184$ kJ/mol $= \pm 1$ kcal/mol.

4. CONCLUSIONS

In the present study, we have assessed the performance of free energy calculations for drug design for a total of 92 ligands binding to five different target proteins. Automating the setup procedure was pivotal to make this large assessment possible and to allow the future use of free energy calculations in drug design projects.

We found that in some cases errors due to insufficient sampling were substantial as indicated by deviating results obtained from two independent runs starting from different ligand coordinates. On average results obtained from the two independent TI runs differed by 3.1 kJ/mol (RMSE). By comparison differences between the TI and MBAR estimators were found to be negligible (RMSE = 0.7 kJ/mol). Analytical uncertainty estimates obtained from a single free energy simulation were found to be much smaller than the observed sample standard deviation obtained from two independent free energy calculations. Interestingly, no increase of the error with increasing size of the perturbation was found. Furthermore, no influence of the absolute size of the single perturbation ($|\Delta G|$) on the precision and accuracy of the free energy differences was observed.

The agreement of calculated with experimental binding affinity differences was found to be system dependent. The accuracy varied from excellent as obtained for the MUP-I system (RMSE = 0.9 kJ/mol, $r = 0.98$) to mediocre for in-house system three (RMSE = 8.7 kJ/mol, $r = 0.5$). For all systems, agreement with experiment improved when restricting analyses to free energy calculations with sample standard deviations below 1 kJ/mol (RMSE = [0.8, 6.9, 1.8, 3.9, 5.6] kJ/mol vs RMSE = [0.9, 8.2, 4.7, 5.7, 8.7] kJ/mol). Analysis of deviations by perturbation type identified most likely suboptimal parametrizations of certain heterocyclic systems and functional groups such as isoxazole and molecules containing sulfur or bromine.

dx.doi.org/10.1021/ci4004199 | *J. Chem. Inf. Model.* 2014, 54, 108−120

Technically speaking, calculation of binding affinity differences for drug design purposes is feasible. Large scale application has so far been hampered by the tedious setup of these calculations and the difficulty to estimate their accuracy. These two challenges have been addressed in the present study. Automation makes larger scale applications possible and enables delivery times (i.e., setup plus runtime) that can compete with experiment. Accuracy estimation remains, however, challenging.

The convergence checks described in this work could be further expanded as follows. Instead of starting from two slightly different binding mode models, several conceivable binding modes could be used as starting configurations of independent free energy simulations to obtain a better estimate of the sample standard deviation. Furthermore, additional perturbations could be introduced (i.e., moving from a spanning tree of perturbations to a graph with cycles) to test whether the free energy difference along a closed cycle is zero. Sampling could be improved by using enhanced sampling methods to speed up the collection of uncorrelated samples.[55−57] Usage of these techniques should be more productive than a simple prolongation of simulation time. We hope that the coming years will bring more large scale applications of free energy calculations which will deliver better statistics on problematic ligand parametrizations and help to improve current general force fields.

■ ASSOCIATED CONTENT

Supporting Information

Flowchart showing the automatic setup procedure (Figure S1), Example input file for `tleap` (Listing S1), calculated and experimental free energy differences for the PDE5 system (Table S1), $\langle \partial V(\lambda)/\partial\lambda \rangle$ vs λ for perturbation 11 → 13 of the PDE5 system (Figure S2), $\langle \Delta\Delta G^{TI}_{c-1} \rangle$ vs production simulation time for perturbation 11 → 13 of the PDE5 system (Figure S3), calculated and experimental free energy differences for in-house systems I−III (Tables S2−S4) This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: thomas.fox@boehringer-ingelheim.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Sandra Handschuh for preparing binding modes for in-house system II. Helpful discussions with David Mobley and Daniel Seeliger are greatfully acknowledged. The authors thank Thomas Steinbrecher for helpful discussions, resolving SHAKE problems in softcore thermodynamic integration simulations, and for providing a script to extract potential energy values from AMBER output for MBAR analysis. The authors thank Hannes G. Wallnoefer for establishing the first steps towards automated TI calculations in our group.

■ ABBREVIATIONS

TI, thermodynamic integration; MBAR, multistate Bennett acceptance ratio; MCS, maximum common substructure; NVT, denotes simulations in the NVT ensemble, i.e. at constant number of particles, volume, and temperature; NPT, denotes simulations in the NPT ensemble, i.e. at constant number of particles, pressure, and temperature

■ REFERENCES

(1) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining electronic laboratory notebooks: Analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745−1756.

(2) Michel, J.; Foloppe, N.; Essex, J. W. Rigorous free energy calculations in structure-based drug design. *Mol. Inf.* **2010**, *29*, 570−578.

(3) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: Progress and challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150−160.

(4) Shirts, M. R. Best practices in free energy calculations for drug design. In *Computational Drug Discovery and Design*; Baron, R., Ed.; 2012; Vol. *819*, Chapter 26, pp 425−467.

(5) Baron, R.; McCammon, J. A. Molecular Recognition and Ligand Association. *Annu. Rev. Phys. Chem.* **2013**, *64*, 151−175.

(6) Skillman, A. G. SAMPL3: Blinded prediction of host-guest binding affinities, hydration free energies, and trypsin inhibitors. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 473−474.

(7) Muddana, H. S.; Varnado, C. D.; Bielawski, C. W.; Urbach, A. R.; Isaacs, L.; Geballe, M. T.; Gilson, M. K. Blind prediction of host-guest binding affinities: A new SAMPL3 challenge. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 475−487.

(8) Timm, D. E.; Baker, L. J.; Mueller, H.; Zidek, L.; Novotny, M. V. Structural basis of pheromone binding to mouse major urinary protein (MUP-I). *Protein Sci.* **2001**, *10*, 997−1004.

(9) Labute, P. *Protonate 3D*, 2007. http://www.chemcomp.com/journal/proton.htm (accessed February 6, 2013).

(10) Sharrow, S. D.; Novotny, M. V.; Stone, M. J. Thermodynamic analysis of binding between mouse major urinary protein-I and the pheromone 2-sec-butyl-4,5-dihydrothiazole. *Biochemistry* **2003**, *42*, 6302−6309.

(11) Haning, H.; Niewhner, U.; Schenke, T.; Lampe, T.; Hillisch, A.; Bischoff, E. Comparison of different heterocyclic scaffolds as substrate analog PDE5 inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3900−3907.

(12) Zhang, K. Y. J.; Card, G. L.; Suzuki, Y.; Artis, D. R.; Fong, D.; Gillette, S.; Hsieh, D.; Neiman, J.; West, B. L.; Zhang, C.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Bollag, G. A glutamine switch mechanism for nucleotide selectivity by phosphodiesterases. *Mol. Cell* **2004**, *15*, 279−286.

(13) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(14) Case, D. A.; Cheatham, T. E., III; 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.

(15) Case, D.; Darden, T.; Cheatham, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Crowley, M.; Walker, R.; Zhang, W.; Merz, K.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossovary, I.; Wong, K.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D.; Seetin, M.; Sagui, C.; Babin, V.; Kollman, P. *Amber 11*; University of California, San Fransisco, 2010.

(16) *OEChem*, version 1.7.2.4; OpenEye Scientific Software Inc.: Santa Fe, NM, 2009.

(17) PDB. http://www.rcsb.org/pdb/ (accessed February 8, 2013).

(18) Dalby, A.; Nourse, J. G.; Douglas Hounshell, W.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(19) Pearlman, D. A. A Comparison of Alternative Approaches to Free Energy Calculations. *J. Phys. Chem.* **1994**, *98*, 1487−1493.
(20) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* **2007**, *127*, No. 214108.
(21) Amber 11. http://ambermd.org/doc11/Amber11.pdf (accessed February 4, 2013).
(22) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(23) Amber Tools. http://ambermd.org/doc11/AmberTools.pdf (accessed February 4, 2013).
(24) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(25) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(26) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(27) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Genet.* **2006**, *65*, 712−725.
(28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(29) Smith, D. E.; Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **1994**, *100*, 3757−3766.
(30) Aaqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021−8024.
(31) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(32) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(33) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(34) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides: The frictional dependence of isomerization rates of *N*-acetylalanyl-*N*′-methylamide. *Biopolymers* **1992**, *32*, 523−535.
(35) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad seeds sprout perilous dynamics: Stochastic thermostat induced trajectory synchronization in biomolecules. *J. Chem. Theory Comput.* **2009**, *5*, 1624−1631.
(36) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation. *J. Chem. Phys.* **2004**, *120*, 6363−6374.
(37) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(38) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(39) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, No. 124105.
(40) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245−268.
(41) 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.
(42) Paliwal, H.; Shirts, M. R. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J. Chem. Theory Comput.* **2011**, *7*, 4115−4134.
(43) Shirts, M.; Chodera, J. https://simtk.org/home/pymbar (accessed February 6, 2013).
(44) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer: Berlin Heidelberg, 2008; pp 319−326.
(45) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0; http://www.R-project.org .
(46) Mobley, D. L. Let's get honest about sampling. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 93−95.
(47) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct calculation of the binding free energies of FKBP ligands. *J. Chem. Phys.* **2005**, *123*, No. 084108.
(48) Zagrovic, B.; van Gunsteren, W. F. Computational Analysis of the Mechanism and Thermodynamics of Inhibition of Phosphodiesterase 5A by Synthetic Ligands. *J. Chem. Theory Comput.* **2007**, *3*, 301−311.
(49) Hansen, N.; Hünenberger, P. H.; van Gunsteren, W. F. Efficient Combination of Environment Change and Alchemical Perturbation within the Enveloping Distribution Sampling (EDS) Scheme: Twin-System EDS and Application to the Determination of OctanolWater Partition Coefficients. *J. Chem. Theory Comput.* **2013**, 1334−1346.
(50) Wang, L.; Berne, B. J.; Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 1326−1330.
(51) Kastenholz, M. A.; Hünenberger, P. H. Computation of methodology-independent ionic solvation free energies from molecular simulations. II. the hydration free energy of the sodium cation. *J. Chem. Phys.* **2006**, *124*, No. 224501.
(52) Reif, M. M.; Hünenberger, P. H. Computation of methodology-independent single-ion solvation properties from molecular simulations. III. Correction terms for the solvation free energies, enthalpies, entropies, heat capacities, volumes, compressibilities, and expansivities of solvated ions. *J. Chem. Phys.* **2011**, *134*, No. 144103.
(53) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the prediction of absolute solvation free energies using the next generation opls force field. *J. Chem. Theory Comput.* **2012**, *8*, 2553−2558.
(54) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data - A Statistical Analysis. *PLoS ONE* **2013**, *8*, e61007.
(55) Jiang, W.; Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2010**, *6*, 2559−2565.
(56) Lapelosa, M.; Gallicchio, E.; Levy, R. M. Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 47−60.
(57) Wang, L.; Berne, B. J.; Friesner, R. A. On achieving high accuracy and reliability in the calculation of relative proteinligand binding binding affinities. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 1937−1942.
(58) Zidek, L.; Stone, M. J.; Lato, S. M.; Pagel, M. D.; Miao, Z.; Ellington, A. D.; Novotny, M. V. NMR mapping of the recombinant mouse major urinary protein I binding site occupied by the pheromone 2-sec-buty1−4,5-dihydrothiazole. *Biochemistry* **1999**, *38*, 9850−9861.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on December 19, 2013, with an incorrect Supporting Information. The corrected version was reposted January 7, 2014.